

ĐẠI HỌC HUẾ
TRƯỜNG ĐẠI HỌC NÔNG LÂM

TS. LÊ ĐÌNH PHÙNG

GIÁO TRÌNH
**PHƯƠNG PHÁP THÍ NGHIỆM
TRONG CHĂN NUÔI VÀ THÚ Y**
(Dùng cho học viên và sinh viên ngành Chăn nuôi - Thú y)

NHÀ XUẤT BẢN NÔNG NGHIỆP
HÀ NỘI - 2010

MỤC LỤC

Lời nói đầu	7
Chương 1. MỘT SỐ KHÁI NIỆM TRONG THỐNG KÊ VÀ PHƯƠNG PHÁP THÍ NGHIỆM	13
1.1. Biến và phân loại biến	13
1.2. Tham số	15
1.3. Thống kê mô tả và thống kê suy diễn	16
1.4. Quần thể và mẫu	17
1.5. Thí nghiệm và nghiên cứu quan trắc	17
1.6. Vật liệu thí nghiệm và đơn vị thí nghiệm	18
1.7. Nhân tố thí nghiệm và nghiệm thức	19
1.8. Lập lại và nhắc lại	19
1.9. Đối chứng	20
1.10. Quy tắc 3R trong thiết kế thí nghiệm	20
1.11. Khối và tạo khối	21
1.12. Sai số thí nghiệm	21
1.13. Phân loại nghiên cứu	22
1.14. Tiến trình nghiên cứu	24
1.15. Bài tập	26
Chương 2. THỐNG KÊ MÔ TẢ	28
2.1. Các tham số (thống kê) đặc trưng cho mức độ tập trung	28
2.1.1. Trung bình cộng	28
2.1.2. Trung bình nhân	29
2.1.3. Trung vị	30
2.1.4. Yếu số	30
2.1.5. Trung bình lược giản	32
2.1.6. Trung bình ba phần	32
2.2. Các tham số (thống kê) đặc trưng cho mức độ phân tán	33
2.2.1. Khoảng biến thiên	33
2.2.2. Khoảng biến thiên 50% ở giữa	33
2.2.3. Phương sai	33
2.2.4. Độ lệch chuẩn	35
2.2.5. Hệ số biến dị	36

2.2.6. Sai số của số trung bình	36
2.2.7. Sự chuyển dịch tuyến tính	37
2.3. Hình dạng phân phối	38
2.4. Trình bày các quan trắc bằng biểu đồ	39
2.4.1. Tổ chức đồ	39
2.4.2. Biểu đồ phân tán	40
2.4.3. Biểu đồ dạng thanh	40
2.4.4. Biểu đồ dạng bánh	41
2.4.5. Biểu đồ dạng cành và lá	41
2.4.6. Đồ thị dạng hộp	42
2.4.7. Biểu đồ xác suất tích lũy	42
2.5. Bài tập	44
Chương 3. KIỂM TRA GIẢ THUYẾT VÀ LỰC THỐNG KÊ	46
3.1. Kiểm tra giả thuyết	46
3.1.1. Khái niệm và phân loại giả thuyết	46
3.1.2. Phương pháp tiếp cận trong kiểm tra giả thuyết	47
3.1.3. Sai sót loại I và mức ý nghĩa	48
3.1.4. Sai sót loại II và lực thống kê của phép kiểm tra	51
3.2. Phân tích lực thống kê và ví dụ về phân tích lực thống kê	52
3.2.1. Mối quan hệ của các yếu tố với lực thống kê	52
3.2.2. Một số ví dụ về lực thống kê	54
3.3. Dung lượng mẫu cần thiết	56
3.3.1. Dung lượng mẫu trong kiểm tra sự khác nhau giữa 2 trung bình nghiệm thức	57
3.3.2. Dung lượng mẫu trong kiểm tra sự khác nhau giữa trung bình các nghiệm thức	59
3.3.3. Dung lượng mẫu cần thiết cho các số liệu phân thành hai nhóm	62
3.4. Bài tập	64
Chương 4. MÔ HÌNH THỐNG KÊ VÀ PHÂN TÍCH PHƯƠNG SAI	65
4.1. Mô hình thống kê	65
4.1.1. Mô hình thống kê cho thí nghiệm	65
4.1.2. Mô hình đầy đủ và mô hình rút gọn và giả thuyết nghiên cứu	66
4.2. Phân tích phương sai	67
4.2.1. Lý do phân tích phương sai và điều kiện để phân tích phương sai	67
4.2.2. Các bước phân tích phương sai	68
4.3. Đa so sánh	76

4.3.1. Giả thuyết bổ sung và hệ số cân đối	77
4.3.2. So sánh có kế hoạch	79
4.3.3. So sánh không có kế hoạch và không độc lập (Kiểm tra Scheffé)	84
4.3.4. So sánh có kế hoạch và không độc lập (Kiểm tra Dunnett')	86
4.3.5. So sánh theo cặp của tất cả các nghiệm thức	88
4.4. Bài tập	92
Chương 5. PHÂN TÍCH PHƯƠNG SAI HAI NHÂN TỐ	94
5.1. Hai thí nghiệm một nhân tố hay một thí nghiệm hai nhân tố ?	94
5.2. Cơ sở của phân tích phương sai hai nhân tố	95
5.2.1. Các thành phần của phương sai	95
5.2.2. Giả thuyết H_0 và H_A	97
5.2.3. Kiểm tra giả thuyết H_0	97
5.3. Tiến hành phân tích phương sai hai nhân tố, một ví dụ thực tế	98
5.4. Các mô hình phân tích phương sai hai nhân tố	105
5.5. Đa so sánh trong ANOVA hai nhân tố	106
5.5.1. Các so sánh có kế hoạch	107
5.5.2. Kiểm tra Scheffé	109
5.5.3. Kiểm tra Tukey HSD	109
5.5.4. Kiểm tra Fisher LSD	110
5.6. Phân tích phương sai hai nhân tố với số lần lặp lại không giống nhau	110
5.6.1. Lặp lại theo tỷ lệ	110
5.6.2. Lặp lại không theo tỷ lệ	112
5.6.3. Phân tích phương sai hai nhân tố không có lần lặp lại	113
5.7. Bài tập	115
Chương 6. THIẾT KẾ THÍ NGHIỆM MỘT NHÂN TỐ	117
6.1. Thiết kế thí nghiệm ngẫu nhiên hoàn toàn	117
6.1.1. Định nghĩa	117
6.1.2. Mô tả thiết kế thí nghiệm ngẫu nhiên hoàn toàn	117
6.1.3. Ngẫu nhiên hóa	119
6.1.4. Mô hình thống kê và phân tích số liệu	122
6.2. Thiết kế thí nghiệm ngẫu nhiên theo khối đầy đủ	125
6.2.1. Định nghĩa	125
6.2.2. Mô tả thiết kế thí nghiệm ngẫu nhiên theo khối đầy đủ	126
6.2.3. Ngẫu nhiên hóa	127
6.2.4. Mô hình thống kê và phân tích số liệu	128

6.3. Thiết kế thí nghiệm hình vuông la tinh	136
6.3.1. Định nghĩa	136
6.3.2. Mô tả thiết kế thí nghiệm hình vuông la tinh	136
6.3.3. Phương pháp ngẫu nhiên hóa trong thiết kế thí nghiệm kiểu ô vuông la tinh	138
6.3.4. Mô hình thống kê và phân tích số liệu	140
6.3.5. Thiết kế thí nghiệm đa ô vuông la tinh	144
6.4. Bài tập	148
Chương 7. PHÂN TÍCH KHÍ BÌNH PHƯƠNG (χ^2)	151
7.1. Phân bố khí bình phương	151
7.2. Kiểm tra sự phân bố lý thuyết	154
7.3. Kiểm tra tính độc lập	156
7.4. Kiểm định chính xác Fisher đối với bảng tương liên 2×2	160
7.5. Điều kiện sử dụng kiểm tra khí bình phương và kiểm tra chính xác Fisher	162
7.6. Bài tập	163
Chương 8. PHÂN TÍCH TƯƠNG QUAN HỒI QUY	164
8.1. Tương quan	164
8.1.1. Biểu đồ phân tán và hiệp phương sai	164
8.1.2. Hệ số tương quan Pearson	166
8.1.3. Kiểm tra giả thuyết về hệ số tương quan Pearson	171
8.2. Phân tích hồi quy	173
8.2.1. Giới thiệu phương trình hồi quy	173
8.2.2. Tiếp cận ước tính hệ số hồi quy	175
8.2.3. Ước tính b và a của phương trình hồi quy	177
8.2.4. Khoảng tin cậy của giá trị ước tính của Y	181
8.2.5. Các thành phần phương sai và hệ số quyết định	182
8.2.6. Kiểm tra giả thuyết phương trình hồi quy và các hệ số của phương trình hồi quy	184
8.3. Bài tập	186
Chương 9. NGHIÊN CỨU ĐIỀU TRA BẰNG BẢN HỎI, BẢN KIỂM	194
9.1. Khái niệm	194
9.2. Vị trí của nghiên cứu điều tra bằng bản hỏi trong hệ thống các nghiên cứu	195
9.3. Các bước trong tiến hành nghiên cứu điều tra bằng bản hỏi	196
9.3.1. Xác định mục tiêu nghiên cứu	197

9.3.2. Xác định giả thuyết nghiên cứu	197
9.3.3. Xác định nhân tố nghiên cứu	197
9.3.4. Xác định chỉ số nghiên cứu	198
9.3.5. Lượng hóa chỉ số nghiên cứu	199
9.3.6. Xây dựng mô hình định tính	199
9.3.7. Phát triển bản hỏi	202
9.3.8. Thử nghiệm bản hỏi	204
9.3.9. Điều tra chính thức	205
9.3.10. Nhập số liệu	207
9.3.11. Phân tích số liệu	209
9.4. Bài tập	211
 Phụ lục 1. Bảng phân bố chuẩn hóa (Z)	 212
Phụ lục 2. Bảng phân bố Student's t	214
Phụ lục 3. Bảng phân bố khi bình phương	216
Phụ lục 4. Bảng phân bố Fisher	218
Phụ lục 5. Giá trị quyết định q với mức $\alpha = 0,05$ (hàng trên) và $\alpha = 0,01$ (hàng dưới) trong kiểm tra Tukey's HSD	229
Phụ lục 6. Giá trị quyết định của hệ số tương quan Pearson	230
Phụ lục 7. Giá trị quyết định của kiểm tra Dunnette	231
Phụ lục 8. Đường cong xác định dung lượng mẫu trong mô hình cố định (v_1 độ tự do tử số, v_2 độ tự do mẫu số)	232
Phụ lục 9. Ví dụ về bản hỏi/bản kiểm điều tra khả năng sinh sản của bò Vàng và bò lai (Vàng $\times$ Brahman) ở vùng sinh thái đồng bằng và miền núi, tỉnh Bình Định	236
Phụ lục 10: Ví dụ về bản hỏi/bản kiểm điều tra ảnh hưởng của giống và vùng sinh thái đến khả năng và hiệu quả chăn nuôi bò ở tỉnh Bình Định	237
 TÀI LIỆU THAM KHẢO CHÍNH	 242

LỜI NÓI ĐẦU

“Đừng hỏi tôi đã học những kiến thức cụ thể gì, hãy hỏi tôi phương pháp tiếp cận nào mà tôi đã học được” (Le, 2006). Điều này hàm ý rằng phương pháp nghiên cứu đóng vai trò quan trọng trong nghiên cứu khoa học. Có được phương pháp nghiên cứu hợp lý sẽ giúp cho nhà nghiên cứu trả lời được câu hỏi nghiên cứu một cách chính xác và hiệu quả nhất. Phương pháp nghiên cứu chính là công cụ biến câu hỏi nghiên cứu, giả thuyết nghiên cứu thành kết luận.

Sau một thời gian học tập và giảng dạy về lĩnh vực phương pháp thí nghiệm trong chăn nuôi và thú y, tôi đã mạnh dạn biên soạn giáo trình này để giúp cho các nhà nghiên cứu, các học viên cao học và sinh viên đại học chuyên ngành chăn nuôi và thú y có thể tham khảo.

Để có thể tiếp thu nội dung của giáo trình này, người đọc cần có một số kiến thức cơ bản về xác suất và thống kê. Cuốn giáo trình này được biên soạn dựa trên nguyên tắc cốt lõi là giúp cho người đọc hiểu bản chất của vấn đề hơn là các bước cụ thể để tính toán các tham số thống kê hay diễn dịch thống kê. Giáo trình này gồm có 9 chương.

Chương 1: Một số khái niệm cơ bản trong thống kê và thiết kế thí nghiệm

Chương 2: Thống kê mô tả

Chương 3: Kiểm tra giả thuyết và lực thống kê

Chương 4: Mô hình thống kê và phân tích phương sai

Chương 5: Phân tích phương sai hai nhân tố

Chương 6: Thiết kế thí nghiệm một nhân tố

Chương 7: Phân tích khi bình phương

Chương 8: Phân tích tương quan hồi quy

Chương 9: Nghiên cứu điều tra bằng bản hỏi, bản kiểm

Để hoàn thành giáo trình này, tôi xin trân trọng cảm ơn PGS.TS. Lê Đức Ngoan đã phản biện cho giáo trình. Các ý kiến phản biện đã góp phần nâng cao chất lượng của giáo trình.

Mặc dù đã cố gắng rất nhiều, giáo trình này không thể không tránh khỏi những sai sót nhất định, tôi kính mong nhận được nhiều ý kiến đóng góp của bạn đọc nói chung, Quý Thầy Cô giáo, học viên cao học, sinh viên đại học chuyên ngành chăn nuôi và thú y nói riêng để nâng cao chất lượng cho các lần tái bản tiếp theo. Mọi góp ý xin được gửi đến email: phung.ledinh@gmail.com.

Xin trân trọng cảm ơn!

Huế năm 2010
TS. Lê Đình Phùng

MỘT SỐ KÝ HIỆU VÀ THUẬT NGỮ THÔNG DỤNG

Ký hiệu	Nghĩa
$\hat{y}$	Giá trị ước tính của y
$\sum$	Tổng
$\sum_{i=1}^n x_i$	Tổng các giá trị
$\sum_{i=1}^n x_i^2$	Tổng các bình phương của các giá trị
$\left(\sum_{i=1}^n x_i\right)^2$	Tổng bình phương của giá trị
$\sum_{i=1}^n xy$	Tổng các tích giữa các giá trị của hai biến xác định trên cùng một đơn vị thí nghiệm
$\bar{X}$ hay M	Giá trị trung bình của mẫu
$1 - \beta$	Lực thống kê
a	Tung độ gốc của đường hồi quy tuyến tính
Adj R ²	Hệ số tương quan bội đã hiệu chỉnh (hệ số quyết định đã hiệu chỉnh)
ANOVA (Analyses of Variance)	Phân tích phương sai
b	Hệ số gốc của đường hồi quy tuyến tính
CI	Khoảng tin cậy
Confounded effect	Ảnh hưởng không tách bạch
Contingency table	Bảng tương liên
Correlation	Tương quan
Covariance	Hiệp phương sai
CRD (Completely Randomized Design)	Thiết kế thí nghiệm hoàn toàn ngẫu nhiên
CV%	Hệ số biến dị (Hệ số biến động)
d	Giá trị chênh lệch giữa 2 giá trị
Dependent variable	Biến phụ thuộc
df (Degree of Freedom)	Độ tự do
Distribution function	Hàm phân phối
F	Phân bố Fisher
f (frequency)	Tần suất
Geomean	Trung bình nhân

Ký hiệu	Nghĩa
H_0 (Null Hypothesis)	Giả thuyết H_0
H_A hay H_1 (Alternative Hypothesis)	Đối thuyết
Independent variable	Biến độc lập
Intercept	Tung độ góc/hằng số
IQR (Inter quartle Range)	Khoảng biến động 50% ở giữa
k hay t	Số nghiệm thức
Kurtosis	Độ tù của phân bố
Linear regression	Hồi quy tuyến tính
LSD (Latin Square Design)	Thiết kế thí nghiệm hình vuông la tinh
LSD (Least Significant Difference)	Sai khác ít nhất có ý nghĩa
Median	Trung vị
Mode	Yếu số
MS_B (Mean Square Between Treatments)	Trung bình bình phương giữa các nhóm hay còn gọi là trung bình bình phương nghiệm thức
MS_E (Mean Square of Error)	Trung bình bình phương của ngẫu nhiên (sai số) = Phương sai của sai số
Non - linear regression	Hồi quy phi tuyến tính
Non - parameter test	Phân tích phi tham số
Normal distribution/Gaussian distribution	Phân bố chuẩn
One tailed test	Kiểm tra một đuôi (một phía)
Parameters	Tham số quần thể
<i>Post hoc</i>	Phân tích sau phương sai
Probability	Xác suất
Probability density function	Hàm mật độ xác suất
Q_1 hay $Q_{25\%}$	Phân vị thứ 25%
Q_2 hay $Q_{50\%}$	Phân vị thứ 50%
Q_3 hay $Q_{75\%}$	Phân vị thứ 75%
Qualitative variable	Biến định tính
Quantitative variable	Biến định lượng
Quartile	Tứ phân vị
r	Hệ số tương quan tuyến tính xác định trên mẫu
R (Range)	Khoảng biến động
r^2	Hệ số xác định
R^2	Hệ số tương quan bội
Rand()	Hàm ngẫu nhiên các giá trị trong khoảng [0 - 1]
Randbetween ()	Hàm ngẫu nhiên trong khoảng giá trị nhất định
Random	Ngẫu nhiên
RCB (Randomized Complete Block Design)	Thiết kế thí nghiệm khối ngẫu nhiên đầy đủ
s	Độ lệch chuẩn của mẫu

Ký hiệu	Nghĩa
s^2	Phương sai mẫu
Sample size	Kích thước mẫu
s_e	Sai số chuẩn
Skew	Độ nhọn của phân bố
Slope	Hệ số góc
SS (Sum of Square)	Tổng các bình phương các độ lệch giữa các quan trắc với giá trị trung bình (còn gọi là tổng các bình phương)
SS_B (Sum of Square Between Treatments)	Tổng các bình phương giữa các nhóm hay giữa các nghiệm thức
SS_T (Total Sum of Square)	Tổng các bình phương tổng số
SS_W (Sum of Square Within each Treatment)	Tổng các bình phương trong các nhóm
Standard deviation	Độ lệch chuẩn
Statistics	Thống kê mẫu (Tham số mẫu)
t	Phân bố student
Trimean	Trung bình ba phần
Two tailed test	Kiểm tra hai đuôi (hai phía)
Variable	Biến
z	Giá trị chuẩn hóa
α	Sai sót loại I
β	Sai sót loại II
μ	Trung bình quần thể
ρ	Hệ số tương quan tuyến tính của quần thể
σ	Độ lệch chuẩn của quần thể
σ^2	Phương sai quần thể
χ^2	Phân bố khi bình phương
Significant level	Mức ý nghĩa
Box plot	Biểu đồ dạng hộp
Stem and leaf	Biểu đồ dạng cành và lá
Histogram	Tổ chức đồ
Scatter plot	Biểu đồ phân tán
Power of the test	Lực thống kê
Contrast	Hệ số tương phản
Planned comparison	So sánh có kế hoạch
Orthogonal comparison	So sánh trực giao
Pairwise comparison	So sánh theo cặp
f_e	Tần suất mong đợi
f_o	Tần suất quan sát

Chương 1

MỘT SỐ KHÁI NIỆM TRONG THỐNG KÊ VÀ PHƯƠNG PHÁP THÍ NGHIỆM

Chương này đề cập đến các khái niệm cơ bản trong thống kê và thiết kế thí nghiệm như biến, phân loại biến, đơn vị thí nghiệm, số lần lặp lại, nhân tố thí nghiệm... Song song với trình bày lý thuyết là các ví dụ được đưa ra để minh chứng cho các khái niệm.

1.1. Biến và phân loại biến

1.1.1. Khái niệm

Biến (variable) là đặc tính của các cá thể, đối tượng nghiên cứu hoặc các sự kiện mà chúng có thể nhận các giá trị khác nhau. Ví dụ các biến có thể là chiều dài thân chéo của bò (cm), vòng ngực của bò (cm), tăng trọng hàng ngày của lợn (g/ngày), giới tính của gia cầm (trống/mái). Giá trị của biến có thể thay đổi từ cá thể (có thể khái quát hóa là đơn vị thí nghiệm) này sang cá thể khác.

1.1.2. Phân loại

Có nhiều cách phân loại biến khác nhau như biến định tính và biến định lượng; biến liên tục và biến rời rạc; biến độc lập và biến phụ thuộc...

1.1.2.1. Biến định tính và biến định lượng

- Biến định tính (qualitative variable): là biến mà giá trị của nó khác nhau về chủng loại; biến định tính có thể bao gồm biến định danh (nominal) và biến thứ hạng (ranked). Biến định danh còn có thể được gọi là biến phân tử (category). Biến định danh có thể bao gồm biến có hai trạng thái và biến gồm nhiều trạng thái. Ví dụ biến giới tính gồm hai trạng thái: đực và cái; biến thụ thai gồm có hai trạng thái: có thụ thai và không có thụ thai. Biến định danh gồm nhiều trạng thái như biến giống lợn nái đang được nuôi ở vùng miền Trung có thể bao gồm: Móng Cái, F1 (Móng Cái $\times$ Yorkshire), Landrace, Yorkshire... Biến định danh có thể được gán một cách tùy ý.

Ví dụ chúng ta có thể mã hóa 1 = giống lợn Móng Cái; 2 = giống lợn Đại Bạch; 3 = giống lợn Landrace. Tính toán các tham số như trung bình hay độ lệch chuẩn không có ý nghĩa đối với dạng biến này.

Một dạng khác của biến định tính còn gọi là thang đo thứ hạng, ví dụ trong thú y người ta có thể xác định mức độ cảm nhiễm ký sinh trùng bằng các mức -; -; +; ++; +++; hay

khi nghiên cứu về mức độ ô nhiễm môi trường từ chất thải chăn nuôi lợn chúng ta có thể phân thang đo như sau: Rất ô nhiễm, khá ô nhiễm, không ô nhiễm.

- **Biến định lượng (quantitative variable):** các giá trị của biến là con số, khác nhau về số lượng. Ví dụ: số lợn con sinh ra/lứa, tăng trọng (g/ngày). Biến định lượng được đo lường dưới dạng số thứ tự, khoảng cách hoặc tỷ lệ. Biến định lượng có thể được phân chia thành biến định lượng biến thiên liên tục và biến định lượng biến thiên rời rạc.

Biến rời rạc: là biến định lượng, nhận một số giá trị nhất định và có khoảng cách ở giữa các giá trị của biến. Ví dụ: số lợn con sinh ra/lứa.

Biến liên tục: là biến định lượng và luôn có giá trị ở giữa hai giá trị của biến. Biến liên tục có thể nhận không giới hạn các giá trị khác nhau. Ví dụ: tăng trọng hàng ngày của lợn (g/ngày), lượng ăn vào tự do của lợn (kg/ngày), vòng ngực của bò (cm), số kg sữa/chu kỳ tiết sữa của bò.

1.1.2.2. Biến độc lập và biến phụ thuộc

- **Biến độc lập:** là biến do người làm nghiên cứu không chế; các mức của biến do người làm nghiên cứu quyết định. Người làm nghiên cứu thay đổi các mức của biến độc lập để quan trắc sự thay đổi của biến phụ thuộc.

- **Biến phụ thuộc:** là biến mà giá trị của nó thay đổi khi biến độc lập thay đổi. Giá trị của biến phụ thuộc được xác định trong quá trình thí nghiệm. Ví dụ: trong nghiên cứu ảnh hưởng của các mức protein khác nhau trong khẩu phần đến tăng trọng của lợn lai F1 (Móng Cái × Đại bạch) thì mức protein trong khẩu phần là biến độc lập và tăng trọng là biến phụ thuộc.

Phân biệt chính xác biến độc lập và biến phụ thuộc có ý nghĩa quan trọng khi sử dụng các phần mềm tin học để xử lý số liệu. Các phần mềm chỉ cho kết quả chính xác khi chúng ta khai báo chính xác biến độc lập và biến phụ thuộc. Trong các nghiên cứu thí nghiệm, phân biệt biến độc lập và biến phụ thuộc thường dễ dàng. Tuy nhiên, trong nghiên cứu điều tra sự phân biệt thường khó hơn nhiều, bởi vì một biến có thể là biến phụ thuộc trong bối cảnh này nhưng là biến độc lập trong bối cảnh khác.

1.1.2.3. Biến định danh, biến thứ bậc, biến khoảng cách và biến tỷ lệ

Một quan điểm phân loại biến phổ biến khác là: biến định danh (nominal), biến thứ bậc (ordinal), biến khoảng cách (interval) và biến tỷ lệ (ratio). Quan điểm phân loại này dựa trên cơ sở ứng dụng các phân tích thống kê khác nhau cho từng loại biến.

- **Biến thứ bậc** là biến mà các giá trị của nó được sắp xếp theo nghĩa con số. Con số cao thể hiện giá trị cao và con số thấp thể hiện giá trị thấp. Tuy nhiên, sự khác nhau giữa các giá trị của biến là không giống nhau. Ví dụ: để đánh giá ảnh hưởng của các giải pháp dinh dưỡng đến độ khó chịu của mùi (hedonic tone) từ phân lợn, tiêu chuẩn đánh giá mùi của Cộng hòa Liên bang Đức (VDI, 1998) sử dụng thang đo 6 bậc như sau:

- H = 1 = không thấy khó chịu
- H = 2 = hơi khó chịu
- H = 3 = khó chịu vừa vừa
- H = 4 = khá khó chịu
- H = 5 = rất khó chịu
- H = 6 = cực kỳ khó chịu

Rõ ràng độ khó chịu khi H = 6 lớn hơn so với khi H = 2. Tuy nhiên, sự khác nhau giữa H = 6 so với H = 5 khác với khi H = 2 so với H = 1.

- Biến khoảng cách là biến định lượng. Biến này chú đến sự khác nhau giữa hai giá trị. Tỷ số giữa hai giá trị của biến này là không có ý nghĩa. Ví dụ sự khác nhau giữa nhiệt độ của 2 chuồng nuôi 30°C và 25°C là giống như giữa 25°C và 20°C; tuy nhiên, không thể kết luận nhiệt độ ở chuồng nuôi 30°C là nóng gấp đôi so với nhiệt độ chuồng nuôi ở 15°C. Đối với loại biến này giá trị 0 mang tính quy ước. Khi nhiệt độ = 0 không có nghĩa là không lạnh hoặc không nóng.

Các loại biến khác nhau cho phép sử dụng các tham số khác nhau (bảng 1.1).

Bảng 1.1. Các loại biến và tham số thống kê

	Có thể tính	Định danh	Thứ bậc	Khoảng cách	Tỷ lệ
Tần suất phân bố		Có	Có	Có	Có
Trung vị		Không	Có	Có	Có
Cộng hoặc trừ		Không	Không	Có	Có
Trung bình, độ lệch chuẩn, sai số của số trung bình...		Không	Không	Có	Có
Tỷ lệ		Không	Không	Không	Có

- Biến tỷ lệ là biến khoảng cách, nhưng có thể chứa giá trị 0 một cách tự nhiên. Các biến như chiều dài, cân nặng, hoạt động của enzyme là biến tỷ lệ. Ví dụ pH không phải là một biến tỷ lệ bởi vì pH = 0 không có nghĩa là không có tính axit. Khi nói đến biến tỷ lệ, chúng ta nói đến tỷ lệ giữa 2 quan trắc, điều này không thể xảy ra đối với biến khoảng cách. Ví dụ: tăng trọng của lợn 600 g/ngày là gấp đôi so với 300 g/ngày, do vậy tăng trọng là biến tỷ lệ. Tuy nhiên, nhiệt độ của nước 100°C không có nghĩa là nóng gấp đôi 50°C, bởi vì nhiệt độ không phải là một biến tỷ lệ. pH = 3 không có nghĩa là tính axit gấp đôi so với pH = 6, bởi vì pH không phải là biến tỷ lệ.

1.2. Tham số

Tham số còn được gọi là tham biến hay thông số. Tham số đo lường các giá trị của quần thể một cách định lượng, ví dụ: tham số trung bình được sử dụng để đo lường mức độ tập trung của quần thể. Các ký tự Hy Lạp được dùng để ký hiệu cho các tham số của

quần thể. Các tham số thông thường không biết được một cách chính xác mà được ước tính từ các giá trị thống kê được tính từ mẫu. Một số ký hiệu tham số và thống kê phổ biến được thể hiện ở bảng 1.2.

Bảng 1.2. Một số ký hiệu tham số và thống kê thường gặp

	Tham số	Thống kê
Trung bình	μ	M hay $\bar{X}$
Phương sai	σ^2	s^2
Độ lệch tiêu chuẩn	σ	s
Tỷ lệ	π	p
Hệ số tương quan	ρ	r

Thuật ngữ thống kê được hiểu theo một số nghĩa khác nhau. Thống kê được hiểu là một tập hợp các kỹ thuật và tiến trình phân tích các số liệu, diễn dịch số liệu, biểu diễn số liệu và đưa ra các quyết định về các kết quả phân tích số liệu. Theo nghĩa này thuật ngữ thống kê có nghĩa là môn học thống kê. Thuật ngữ thống kê còn được hiểu theo một nghĩa khác đó là các giá trị đo lường bằng các con số mà các giá trị này được tính từ mẫu. Theo nghĩa này, thống kê được dùng để ước tính các tham số.

1.3. Thống kê mô tả và thống kê suy diễn

- Thống kê mô tả (descriptive statistics) được sử dụng để tóm tắt một tập hợp số liệu theo một cách rõ ràng và dễ hiểu. Ví dụ một nghiên cứu tiến hành theo dõi khả năng sinh sản của lợn nái Landrace trên 3000 lứa đẻ, để cung cấp thông tin về khả năng sinh sản của lợn nái Landrace không thể trình bày kết quả của cả 3000 lứa đẻ một cách riêng lẻ mà chỉ trình bày các thống kê đại diện cho các lứa đẻ nghiên cứu. Có hai hướng tiếp cận cơ bản trong thống kê mô tả: (i) tính các giá trị thống kê như trung bình, trung vị, yếu số, phương sai, độ lệch chuẩn...; (ii) xây dựng các biểu đồ như biểu đồ cành và lá (stem and leaf) hay biểu đồ hộp (box plot). Hướng thứ hai thường được sử dụng để phát hiện khuynh hướng, trong khi đó hướng thứ nhất thường khách quan và chính xác hơn. Hai hướng tiếp cận này có khả năng bổ trợ cho nhau do vậy có thể dùng cả hai hướng trong đúc kết các số liệu quan trắc.

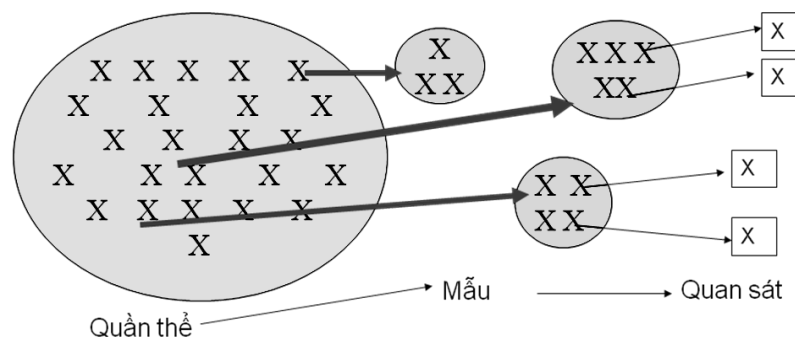
- Thống kê suy diễn (Inferential statistics) được sử dụng để suy diễn kết quả ở quần thể từ các kết quả ở mẫu. Ví dụ: một nhóm nghiên cứu ảnh hưởng của bổ sung hoặc không bổ sung β - glucan vào khẩu phần ăn của lợn con sau cai sữa cho thấy tăng trọng của lợn ở nghiệm thức có bổ sung là 400 g/ngày và đối chứng là 350g/ngày. Câu hỏi đặt ra là liệu sự khác nhau đó là thực sự do ảnh hưởng của bổ sung β - glucan hay là do ngẫu nhiên? Có hai phương pháp chủ yếu trong thống kê suy diễn: ước tính và kiểm

tra giả thuyết. Trong ước tính, thông tin từ mẫu được sử dụng để ước tính tham số và khoảng tin cậy của ước tính. Kiểm tra giả thuyết được sử dụng để quyết định chấp nhận hay bác bỏ giả thuyết H_0 .

1.4. Quần thể và mẫu

Quần thể là tập hợp toàn bộ các đối tượng, quan trắc quan tâm có đặc điểm chung nào đó, ví dụ cùng giống, cùng giới tính, cùng vùng địa lý. Ví dụ quần thể lợn nái Móng Cái ở Thừa Thiên Huế. Phân bố của quần thể có thể được mô tả thông qua các tham số như trung bình và độ lệch chuẩn. Các tham số đó được ước tính từ các thống kê mẫu. Mẫu là một phần của quần thể mà chúng ta quan tâm. Thông thường không thực tế (tốn kém thời gian và kinh phí) để kiểm tra hay đo lường tất cả các đối tượng của quần thể, do vậy mẫu được chọn từ quần thể để nghiên cứu về quần thể. Từ các cá thể được chọn ra từ quần thể, chúng ta đo lường các giá trị (quan trắc) sau đó tiến hành phân tích thống kê để đưa ra kết luận ở mức độ quần thể.

Thông thường sau khi thu thập và phân tích số liệu, các kết luận được đưa ra ở cấp độ quần thể dựa trên các giá trị thống kê của mẫu. Do vậy, các mẫu cần được chọn theo cách có thể đại diện cho quần thể. Thống kê suy diễn thường yêu cầu các mẫu phải được chọn ra một cách ngẫu nhiên.



Biểu đồ 1.1. Quần thể, mẫu và quan trắc

Trong chọn mẫu ngẫu nhiên, mỗi thành phần (đơn vị thí nghiệm) của quần thể có xác suất được chọn như nhau. Một mẫu được gọi là ngẫu nhiên khi phương pháp chọn mẫu đáp ứng tiêu chuẩn ngẫu nhiên. Một mẫu không ngẫu nhiên (mẫu thiên lệch) là mẫu được tạo ra từ phương pháp chọn mẫu không đáp ứng tiêu chuẩn ngẫu nhiên. Mẫu thiên lệch thường sai khác một cách có hệ thống so với quần thể.

1.5. Thí nghiệm và nghiên cứu quan trắc

Một thí nghiệm là nghiên cứu mà thiết lập một tập hợp chi tiết các tình huống dựa trên một quy trình cụ thể để quan trắc, đánh giá các kết quả quan trắc. Nhà nghiên cứu

thiết lập và điều khiển các quy trình trong một thí nghiệm để đánh giá và kiểm tra một hoặc nhiều vấn đề mà chưa có câu trả lời tính đến thời điểm nghiên cứu.

Thí nghiệm so sánh (comparative experiment) là loại thí nghiệm quen thuộc với các nhà nghiên cứu trong lĩnh vực sinh học, dược lý học, nông nghiệp, kỹ thuật, sinh lý. Thí nghiệm so sánh thiết lập hơn một tình huống trong nghiên cứu, phân chia các tình huống đó vào các đối tượng nghiên cứu và quan trắc sự đáp ứng. Đáp ứng của đối tượng nghiên cứu trong các tình huống khác nhau được so sánh với nhau.

Nghiên cứu quan trắc so sánh (comparative observation studies) là loại nghiên cứu mà chúng ta muốn làm thí nghiệm nhưng do điều kiện thực tế hoặc do các yếu tố liên quan đến vật quyền cho nên chúng ta không tiến hành thí nghiệm được. Đối tượng nghiên cứu trong nghiên cứu quan trắc có vai trò như các đơn vị thí nghiệm trong nghiên cứu thí nghiệm. Các đối tượng nghiên cứu tồn tại thành từng nhóm có thể phân biệt được với nhóm khác và mỗi nhóm tồn tại trong các tình huống cụ thể. Thay vì phân chia các đơn vị thí nghiệm vào các nghiệm thức như trong nghiên cứu thí nghiệm, nghiên cứu quan trắc so sánh sử dụng các nhóm và các tình huống mà nó tồn tại. Một ví dụ dễ hiểu không thuộc lĩnh vực chăn nuôi là xem xét ảnh hưởng của việc có hoặc không sử dụng mũ bảo hiểm đến chấn thương sọ não khi tai nạn xe máy xảy ra. Sẽ là không đạo đức để chia đối tượng nghiên cứu thành hai nhóm. Một nhóm đội mũ bảo hiểm và nhóm còn lại không và sau đó cố ý gây tai nạn cho cả hai nhóm và quan trắc sự xuất hiện chấn thương sọ não. Thay vì làm vậy, chúng ta có thể sử dụng các hồ sơ tai nạn xe máy có sẵn và phân ra hai nhóm: có đội mũ bảo hiểm và không đội mũ bảo hiểm.

1.6. Vật liệu thí nghiệm và đơn vị thí nghiệm

Vật liệu thí nghiệm là nguyên liệu để tiến hành thí nghiệm. Đơn vị thí nghiệm là (i) tập hợp nhỏ nhất của vật liệu thí nghiệm, (ii) chịu tác động của yếu tố thí nghiệm và (iii) có thể độc lập với các đơn vị thí nghiệm khác. Thông thường đặc tính thứ 3 của đơn vị thí nghiệm bị lãng quên. Ví dụ một nhóm nghiên cứu ảnh hưởng của 2 khẩu phần ăn A và B đến tăng trọng của lợn. 12 con lợn thí nghiệm được phân chia một cách ngẫu nhiên vào 2 ô chuồng khác nhau, 6 con/ô. Khối lượng của mỗi con lợn được cân định kỳ và riêng lẻ. Một ô được nhận khẩu phần A một cách ngẫu nhiên, ô kia nhận khẩu phần B. Trong ví dụ này, một đơn vị thí nghiệm tương ứng với một ô chuồng có 6 con lợn, mặc dù lợn được cân khối lượng định kỳ riêng lẻ, bởi vì cả 6 con lợn trong một ô chuồng là không độc lập nhau. Một ví dụ khác: một nhóm nghiên cứu quan tâm đến ảnh hưởng của 3 mức protein thô 12; 14 và 16% trong khẩu phần đến tăng trọng của lợn thịt F1 (Landrace × Móng Cái). Thí nghiệm được tiến hành trên 18 con lợn. Mỗi con được nuôi trong một ô chuồng. Trong ví dụ này mỗi một con lợn là một đơn vị thí nghiệm.

Đơn vị thí nghiệm phải mang tính đại diện cho quần thể mà ta cần nghiên cứu. Đơn vị thí nghiệm càng đồng nhất thì sai sót thí nghiệm sẽ càng nhỏ và độ chính xác của thí nghiệm càng lớn. Tuy nhiên, khi đơn vị thí nghiệm quá đồng nhất thì có thể ảnh hưởng đến khả năng ứng dụng vào thực tiễn của kết quả nghiên cứu.

1.7. Nhân tố thí nghiệm và nghiệm thức

Nhân tố thí nghiệm còn gọi là yếu tố thí nghiệm, là nguyên nhân gây ảnh hưởng đến các giá trị quan trắc. Nhân tố thí nghiệm là do người nghiên cứu quyết định. Có thể xem nhân tố thí nghiệm như là biến độc lập. Trong một thí nghiệm có thể có một hoặc nhiều hơn một nhân tố. Nhân tố thí nghiệm có thể là cố định hoặc ngẫu nhiên. Thí nghiệm nhiều nhân tố thường hiệu quả hơn (xét trên góc độ số đơn vị thí nghiệm cần thiết) so với thí nghiệm đơn nhân tố.

Nhân tố nghiên cứu gồm có các mức của nhân tố. Các mức của nhân tố do người nghiên cứu quyết định và được đặt ra trước khi tiến hành nghiên cứu. Mỗi mức của nhân tố còn gọi là mỗi nghiệm thức. Nghiệm thức cũng có thể là sự kết hợp giữa các mức của các nhân tố. Nghiệm thức có thể là định tính cũng có thể là định lượng.

- Nhân tố giống bò có thể có các nghiệm thức: bò Vàng, bò Red Sindhi, bò Brahman.
- Nhân tố mức thức ăn tinh bổ sung vào khẩu phần bò có thể có các nghiệm thức: 0,5; 1; 1,5; 2% tính theo vật chất khô khối lượng cơ thể của bò.
- Kiểu gen Halothan ở lợn có thể có 3 kiểu gen NN, Nn, nn.

Nghiệm thức còn có thể được gọi là công thức thí nghiệm. Sự kết hợp giữa nhân tố giống bò và nhân tố mức thức ăn tinh bổ sung vào khẩu phần như đề cập ở trên tạo nên 12 công thức thí nghiệm.

Để đảm bảo độ chính xác của nghiên cứu, định nghĩa nhân tố nghiên cứu cần phải rõ ràng để không chế tác động của các nhân tố không phải là nhân tố nghiên cứu đến đơn vị thí nghiệm. Định nghĩa nhân tố nghiên cứu thường dễ dàng trong các nghiên cứu thí nghiệm, tuy nhiên trong các nghiên cứu điều tra các nhân tố nghiên cứu thường khó được định nghĩa một cách dễ dàng.

1.8. Lặp lại và nhắc lại

Số lần lặp lại chính là số đơn vị thí nghiệm nhận cùng tác động của một nghiệm thức. Trong thí nghiệm số lần lặp lại có thể giống nhau hoặc khác nhau giữa các nghiệm thức. Để thuận tiện cho phân tích kết quả nghiên cứu, số lần lặp lại nên giống nhau. Tuy nhiên, trong thực tế không phải lúc nào cũng đạt được thí nghiệm có số lần lặp lại bằng nhau, bởi vì đối tượng thí nghiệm (gia súc, gia cầm) có thể bị chết trong thời gian thí nghiệm.

Trong một số trường hợp khi số lần lặp lại không bằng nhau thì phải sử dụng hình thức phân tích số liệu phức tạp như mô hình phân tích hồi quy đa biến hỗn hợp (multiple mixed model); hoặc phải ước tính những giá trị bị mất dựa trên các số liệu đã có trong cùng nghiệm thức.

Khái niệm lặp lại khác với nhắc lại. Nhắc lại là làm lại thí nghiệm trong điều kiện tương tự để kết luận có được độ chính xác cao. Lặp lại thí nghiệm thường được ứng dụng đối với các nghiên cứu trong phòng thí nghiệm cũng như nghiên cứu trên đồng ruộng.

1.9. Đối chứng

Nhóm đối chứng được tạo ra trong quá trình thiết kế thí nghiệm. Nhóm đối chứng và nhóm thí nghiệm được tác động hoàn toàn giống nhau ngoại trừ mức của nhân tố nghiên cứu mà đang được nghiên cứu. Sự thay đổi của biến độc lập dẫn đến sự thay đổi của biến phụ thuộc. Tuy nhiên, nhiều khi sự thay đổi của biến phụ thuộc có thể là kết quả của sự thay đổi của biến ngẫu nhiên. Để đảm bảo hiện tượng này không xảy ra, chúng ta có thể thêm vào thí nghiệm một nhóm đối chứng. Trong thí nghiệm, nhóm đối chứng có thể được xem xét như là một nghiệm thức cung cấp số liệu đối chứng để so sánh với các số liệu thu được từ các nghiệm thức thí nghiệm.

Một ví dụ đơn giản nhất của một thí nghiệm có đối chứng là kiểm tra tác dụng của thuốc trong lĩnh vực thú y. Thí nghiệm chia các đơn vị thí nghiệm ra làm 02 nhóm: nhóm thí nghiệm và nhóm đối chứng. Nhóm được sử dụng thuốc gọi là nhóm thí nghiệm, nhóm không dùng thuốc gọi là nhóm đối chứng.

Một ví dụ khác về ảnh hưởng của bổ sung β - glucan đến sinh trưởng của lợn con sau cai sữa. Để đánh giá được ảnh hưởng của bổ sung β - glucan, thí nghiệm cần có một nhóm đối chứng không được bổ sung β - glucan. Ảnh hưởng của bổ sung β - glucan ở các mức khác nhau đến sinh trưởng của lợn con sau cai sữa có thể được so sánh với nhóm đối chứng.

1.10. Quy tắc 3R trong thiết kế thí nghiệm

Một thí nghiệm yêu cầu phải đảm bảo ít nhất hai đặc tính: tính ngẫu nhiên (randomization) và tính lặp lại (replication). Trong một số trường hợp ví dụ khi các đơn vị thí nghiệm không đồng đều thì đòi hỏi cần có khống chế sự sai khác ban đầu (restriction).

Yêu cầu tính lặp lại để đảm bảo có hơn một đơn vị thí nghiệm được nhận tác động của mỗi nghiệm thức. Nếu không có tính lặp lại thì kết quả thu được từ thí nghiệm chỉ mang tính chất mô tả mà không thể suy diễn thống kê do không thể ước tính được biến động ngẫu nhiên cũng như biến động do nhân tố thí nghiệm gây nên. Lặp lại còn cho phép ước tính được lực thống kê của thí nghiệm.

Ngẫu nhiên trong ngữ cảnh thiết kế thí nghiệm có nghĩa là quá trình phân chia các đơn vị thí nghiệm vào các nghiệm thức hoặc ngược lại phải đảm bảo nguyên tắc ngẫu nhiên. Có nghĩa là mỗi đơn vị thí nghiệm đều có cơ hội như nhau khi nhận tác động của nghiệm thức. Khi thiết kế thí nghiệm cần phải ngẫu nhiên để hạn chế sự thiên lệch. Ngẫu nhiên hóa có thể chuyển các ảnh hưởng mang tính hệ thống và ảnh hưởng của các nhân tố không khống chế được sang dạng ngẫu nhiên.

Khống chế sự sai khác ban đầu được tiến hành khi các đơn vị thí nghiệm không tương đồng khi chuẩn bị bắt đầu thí nghiệm, đặc biệt là không tương đồng về các nhân tố có thể ảnh hưởng đến biến phụ thuộc. Ví dụ: người nông dân A có 12 con bò bị bệnh và có triệu chứng giống nhau, trong đó có 6 con bò 2 năm tuổi và 6 con bò 3 năm tuổi,

người nông dân dùng thuốc cho 3 con 2 năm tuổi và 3 con 3 năm tuổi, sau 1 tuần 6 con hết bệnh, 6 con bò còn lại không được dùng thuốc thì sau 1 tháng mới hết bệnh. Kết quả của thí nghiệm này không tách bạch được ảnh hưởng của việc sử dụng thuốc và ảnh hưởng của sự khác nhau về tuổi. Hiện tượng này còn gọi là hiện tượng đan xen ảnh hưởng (confounded). Sở dĩ xảy ra hiện tượng này là do đã không có sự không chế ban đầu khi các đơn vị thí nghiệm không tương đồng về tuổi. Để nghiên cứu được ảnh hưởng của thuốc cần có sự không chế ban đầu. Một trong những kỹ thuật để không chế ban đầu là sử dụng kỹ thuật khối.

1.11. Khối và tạo khối

Trong thiết kế thí nghiệm, khối là một tập hợp các đơn vị thí nghiệm tương đồng. Các đơn vị thí nghiệm trong một khối tương đồng với nhau nhiều hơn so với các đơn vị thí nghiệm giữa các khối. Tạo khối là sắp xếp các đơn vị thí nghiệm tương đồng vào trong cùng một khối. Thông thường yếu tố tạo khối là một nguồn tạo nên sự biến động của biến phụ thuộc, tuy nhiên ảnh hưởng của nguồn biến động này không được quan tâm bởi các nhà nghiên cứu. Các yếu tố tạo khối có thể là giới tính, khối lượng, tuổi....

Trong thiết kế thí nghiệm, tạo khối làm nâng cao độ chính xác của thí nghiệm thông qua loại bỏ ảnh hưởng của các yếu tố gây nhiễu. Bất kỳ thí nghiệm nào cũng có yếu tố gây nhiễu. Người nghiên cứu cần xác định yếu tố gây nhiễu nào là quan trọng nhất để không chế. Bản chất của không chế yếu tố gây nhiễu thông qua tạo khối là loại bỏ hoặc tách đóng góp của yếu tố gây nhiễu ra khỏi sai số ngẫu nhiên. Nếu không tạo khối thì biến động (có thể dùng thống kê tổng các bình phương các độ lệch) do yếu tố gây nhiễu tạo nên sẽ được bao gồm trong biến động của ngẫu nhiên (SS_E ¹), tuy nhiên khi được tạo khối thì biến động do yếu tố gây nhiễu cụ thể nào đó sẽ được tách ra khỏi biến động của ngẫu nhiên. Tạo khối là tạo ra các nhóm đơn vị thí nghiệm tương đồng, trong các khối như vậy yếu tố gây nhiễu được giữ ổn định, để tạo điều kiện cho nhân tố thí nghiệm biến động. Trong mỗi khối có thể đánh giá ảnh hưởng các mức của nhân tố nghiên cứu mà không phải lo ảnh hưởng không tách bạch của yếu tố gây nhiễu.

Nguyên tắc cơ bản là “Tạo khối những gì có thể và ngẫu nhiên hóa những gì không thể”. Tạo khối dùng để loại bỏ ảnh hưởng của một vài biến nhiễu quan trọng. Sau đó, ngẫu nhiên hóa được sử dụng để giảm các biến nhiễu còn lại.

1.12. Sai số thí nghiệm

Sai số thí nghiệm đo lường sự biến động giữa các đơn vị thí nghiệm mà các đơn vị đó được nhận cùng một nghiệm thức. Trước hết cần phân biệt sai số thí nghiệm (error) và lỗi (mistake). Hai thuật ngữ này không phải là một. Sai số thí nghiệm có nghĩa bao quát hơn so với lỗi. Khi tiến hành một thí nghiệm chúng ta có thể có ba loại sai số sau:

¹ Khái niệm SS_E sẽ được giải thích trong chương 4.

(i) sai số do con người gây nên, (ii) sai số hệ thống và (iii) sai số ngẫu nhiên. Sai số do con người gây nên còn được gọi là lỗi. Sai số do con người gây nên không phải là nguồn gốc của sai số thí nghiệm.

Sai số hệ thống là sai số cố hữu trong thiết kế thí nghiệm. Loại sai số này làm lệch các giá trị đo lường được theo một hướng cố định nào đó, ví dụ luôn luôn lớn hoặc luôn luôn nhỏ hơn giá trị thực của nó. Loại sai số này xuất phát từ sự thiếu đồng nhất trong quá trình thực hiện thí nghiệm. Có thể do dụng cụ đo lường không được hiệu chỉnh một cách thường xuyên. Ví dụ: sử dụng cân để xác định khối lượng, nhưng cân không được điều chỉnh một cách chính xác, do vậy khối lượng của bỏ cân được luôn luôn cao hơn so với thực tế.

Tất cả các thí nghiệm đều có sai số ngẫu nhiên. Loại sai số này xảy ra bởi vì không có đo lường nào có giá trị một cách chính xác hoàn toàn. Sai số ngẫu nhiên tạo nên một tập hợp các quan trắc mà lúc thì quá lớn, lúc thì quá nhỏ hơn so với giá trị thực của nó. Sai số ngẫu nhiên tạo ra biến động không giải thích được.

1.13. Phân loại nghiên cứu

Có nhiều quan điểm phân loại nghiên cứu khác nhau. Mỗi quan điểm phân loại nghiên cứu đều có tính chất tương đối.

- Nghiên cứu điều tra và nghiên cứu thí nghiệm

Nghiên cứu điều tra là một dạng đặc biệt của nghiên cứu quan trắc. Loại nghiên cứu này không có tác động đến đối tượng nghiên cứu mà chỉ mô tả hiện thực về đối tượng nghiên cứu ở trạng thái mà nó đang tồn tại. Có hai loại nghiên cứu điều tra cơ bản: (i) điều tra tại một thời điểm và (ii) điều tra lặp lại. Ví dụ: để đánh giá hiện trạng chăn nuôi bò của một cơ sở sản xuất nào đó trên các tiêu chí như số lượng, quy mô, cơ cấu đàn, cơ cấu giống, tình hình chăm sóc nuôi dưỡng... thì chúng ta có thể sử dụng điều tra tại một thời điểm. Trong khi đó để nghiên cứu sự vận chuyển nitơ trong hệ thống VAC (Vườn, Ao, Chuồng) của nông hộ thì chúng ta cần điều tra nhiều thời điểm để thu thập thông tin về các yếu tố đầu vào, đầu ra của hệ thống, từ đó lượng hóa được dòng chảy nitơ.

Nghiên cứu thí nghiệm là nghiên cứu mà chúng ta có tác động lên đối tượng nghiên cứu, thay đổi điều kiện tồn tại bình thường của đối tượng nghiên cứu để quan trắc sự thay đổi. Cụ thể là áp dụng các nghiệm thức khác nhau lên các đơn vị thí nghiệm và xác định giá trị của các biến phụ thuộc. Một số kiểu thiết kế thí nghiệm cơ bản liên quan đến loại nghiên cứu này bao gồm: ngẫu nhiên hoàn toàn (CRD); khối ngẫu nhiên đầy đủ (RCB); chia lô chia thửa (Split plot) và hình vuông la tinh (LSD).

Giáo trình này dựa trên quan điểm phân loại nghiên cứu này.

- Nghiên cứu cơ bản, ứng dụng và triển khai

Nghiên cứu cơ bản được dùng để phát hiện bản chất của sự vật, hiện tượng, ví dụ như quy luật sinh trưởng và phát triển của cây trồng và vật nuôi, giải mã hệ gen của

virus H₅N₁, hay xác định các loại kháng sinh thực vật trong cây cỏ Xước. Loại nghiên cứu này được tiến hành trong phòng thí nghiệm và đòi hỏi các trang thiết bị và cơ sở vật chất hiện đại cũng như thời gian và kinh phí.

Nghiên cứu ứng dụng là ứng dụng các kết quả của nghiên cứu cơ bản vào trong điều kiện sản xuất thực tế để tìm ra các giải pháp cụ thể ứng dụng vào sản xuất. Ví dụ: biện pháp cai sữa sớm ở lợn con, nghiên cứu thích nghi của một giống cây con nào đó vào hệ thống sản xuất cụ thể, nghiên cứu bổ sung bột cây cỏ Xước vào thức ăn của lợn con sau cai sữa đến tỷ lệ mắc bệnh tiêu chảy và sinh trưởng của lợn con. Nghiên cứu này ứng dụng kết quả của nghiên cứu cơ bản là phát hiện các loại kháng sinh thực vật trong cây cỏ Xước. Loại nghiên cứu này được tiến hành trong điều kiện của trại thí nghiệm. Câu trả lời quan trọng của loại nghiên cứu này là biện pháp kỹ thuật có thể được sử dụng trong một hệ thống sản xuất cụ thể nào đó không.

Nghiên cứu triển khai ứng dụng các kết quả của nghiên cứu cơ bản, nghiên cứu ứng dụng vào trong điều kiện sản xuất thực tế hay nói cách khác là chuyển tải các kết quả nghiên cứu cơ bản, ứng dụng vào điều kiện sản xuất cụ thể. Loại nghiên cứu này được tiến hành trong điều kiện nông hộ. Nghiên cứu này thường so sánh các biện pháp kỹ thuật đang tồn tại với biện pháp kỹ thuật mới từ đó lựa chọn biện pháp phù hợp.

- Nghiên cứu trong phòng thí nghiệm, trạm trại và điều kiện sản xuất

Nghiên cứu trong phòng thí nghiệm là những nghiên cứu được tiến hành trong phòng thí nghiệm, đòi hỏi các thiết bị, hóa chất và đòi hỏi khống chế các yếu tố gây nhiễu cao. Thông thường loại nghiên cứu này là các nghiên cứu cơ bản.

Nghiên cứu ở trạm trại thí nghiệm là nghiên cứu được tiến hành trên đối tượng vật nuôi ở các trạm trại thí nghiệm trong điều kiện thí nghiệm khá giống với thực tế sản xuất. Loại nghiên cứu này thường là nghiên cứu ứng dụng.

Nghiên cứu trong điều kiện sản xuất là nghiên cứu được tiến hành trên cây trồng và vật nuôi, nông hộ trong thực tế sản xuất. Nghiên cứu này thường có quy mô lớn và điều kiện của nghiên cứu là điều kiện thực tế của sản xuất. Nghiên cứu này thường là nghiên cứu triển khai.

- Nghiên cứu định tính và định lượng

Nghiên cứu định tính là nghiên cứu mô tả chất lượng, mô tả bản chất của sự vật hiện tượng. Mục tiêu của nghiên cứu định tính là bối cảnh hóa. Nghiên cứu định tính được tiến hành khi mục tiêu nghiên cứu, đối tượng nghiên cứu, chỉ số nghiên cứu chưa được xác định một cách rõ ràng. Nghiên cứu định tính có độ linh động cao. Các thông tin nghiên cứu thường là từ ngữ hơn là các con số. Các thông tin nghiên cứu được quy nạp để rút ra các quy luật chung.

Nghiên cứu định lượng là nghiên cứu mà thông tin thu thập được là các con số định lượng hoặc các thông tin có thể chuyển tải về các con số. Các thông tin nghiên cứu được diễn dịch thống kê để rút ra kết luận ở cấp độ quần thể trên cơ sở các số liệu từ mẫu.

Nghiên cứu định lượng được tiến hành khi mục tiêu nghiên cứu, đối tượng nghiên cứu và chỉ số nghiên cứu đã được xác định một cách rõ ràng.

Trong quá trình nghiên cứu, nghiên cứu định tính được tiến hành trong giai đoạn đầu và nghiên cứu định lượng được tiến hành trong giai đoạn sau. Nghiên cứu định tính và định lượng không tách rời nhau mà luôn bổ trợ nhau để giải thích kết quả một cách thuyết phục nhất.

Ngoài ra còn có một số quan điểm phân loại khác như:

- Nghiên cứu một nhân tố và nghiên cứu nhiều nhân tố
- Nghiên cứu truyền thống và không truyền thống
- Nghiên cứu mô tả, giải thích, giải pháp và dự báo
- Nghiên cứu quan sát và nghiên cứu thí nghiệm

1.14. Tiến trình nghiên cứu

Nghiên cứu là một quá trình liên kết logic từ hình thành ý tưởng nghiên cứu đến báo cáo kết quả nghiên cứu. Có nhiều cách phân chia tiến trình nghiên cứu khác nhau nhưng nhìn chung xét trên góc độ khoa học thuần túy (xem nhẹ góc độ quản lý và tài chính) thì nghiên cứu có thể có bốn bước cơ bản sau:

- Xác định chủ đề nghiên cứu

Đề tài nghiên cứu cần đáp ứng đầy đủ ba yêu cầu cơ bản: (i) tính khoa học, (ii) tính thực tiễn và (iii) tính cấp thiết. Để xác định đề tài nghiên cứu tác giả cần phải tham khảo nhiều tài liệu khác nhau, đặc biệt là các tạp chí, kỉ yếu hội thảo. Thông thường giáo trình cung cấp các kiến thức cơ bản. Các bài báo trong các tạp chí cung cấp thông tin xác định đề tài nghiên cứu mới. Cần xác định các lỗ hổng nghiên cứu từ các đề tài nghiên cứu đã được tiến hành. Lỗ hổng nghiên cứu (knowledge gap) là cơ sở quan trọng để xác định các đề tài nghiên cứu.

- Xây dựng đề cương nghiên cứu

Sau khi đã xác định đề tài nghiên cứu thì cần xây dựng đề cương nghiên cứu. Có thể xem đề cương nghiên cứu như là một bản thiết kế, bản hướng dẫn để tiến hành nghiên cứu. Đề cương nghiên cứu cần được viết cụ thể đến mức mà người khác có cùng chuyên môn có thể nhắc lại một thí nghiệm. Đề cương nghiên cứu cũng cần chú ý đến tính hấp dẫn, thu hút sự quan tâm của người đọc. Đặc biệt đối với các đề cương nghiên cứu để xin kinh phí nghiên cứu thực hiện các đề tài. Nhìn chung, đề cương nghiên cứu cần có các mục sau đây:

- Tên đề tài nghiên cứu: nên ngắn gọn nhưng phải thể hiện được nội hàm của nghiên cứu. Tên đề tài thường được đọc đầu tiên, trước khi quyết định có tiếp tục xem xét các phần còn lại của đề cương hay báo cáo. Cách đặt tên đề tài nghiên cứu lý tưởng nhất là có sự gắn kết giữa tên đề tài nghiên cứu, mục tiêu nghiên cứu, giả thuyết nghiên cứu, nội dung nghiên cứu và kết luận. Tên đề tài có thể

bao gồm biến độc lập, biến phụ thuộc, đối tượng nghiên cứu, địa điểm nghiên cứu. Ví dụ: “*Ảnh hưởng của các mức protein trong khẩu phần đến khả năng sản xuất của lợn thịt lai (Omega × Yorkshire) nuôi trong điều kiện chăn nuôi công nghiệp*”. Thông thường mỗi cơ quan khoa học có một quy định riêng về tên đề tài nghiên cứu. Ví dụ *Tạp chí chăn nuôi của Mỹ (Journal of Animal Sciences)* quy định tên đề tài không dài quá 12 từ.

- Đặt vấn đề: cần nêu rõ ba yêu cầu của đề tài nghiên cứu. Đặt vấn đề cần trả lời câu hỏi vì sao phải tiến hành đề tài nghiên cứu.
- Mục tiêu nghiên cứu và giả thuyết nghiên cứu: Mục tiêu nghiên cứu có thể nằm ở một phần riêng trong đề cương nghiên cứu nhưng thông thường được nằm ở đoạn văn cuối cùng của phần đặt vấn đề. Cần phân biệt rõ mục tiêu nghiên cứu và mục đích nghiên cứu. Mục đích nghiên cứu trả lời câu hỏi đề tài nghiên cứu đang xem xét để làm gì (for what purpose)? Trong khi đó mục tiêu nghiên cứu chính là hoạt động gì cần làm (for what). Giả thuyết nghiên cứu có thể được xem như là kết luận sơ bộ của đề tài nghiên cứu. Nghiên cứu được tiến hành để đưa ra quyết định chấp nhận hay bác bỏ giả thuyết H_0 .
- Nội dung nghiên cứu: là hoạt động nghiên cứu để đáp ứng các mục tiêu nghiên cứu. Nội dung nghiên cứu cần chỉ rõ các chỉ số nghiên cứu nào.
- Phương pháp nghiên cứu: cần chỉ rõ biến nào là biến độc lập và biến nào là biến phụ thuộc. Cách xác định các biến đó như thế nào? Đối tượng nghiên cứu, cụ thể đơn vị thí nghiệm là gì? Mô tả đầy đủ về đơn vị thí nghiệm, ví dụ giống, tuổi, khối lượng ban đầu,... có bao nhiêu đơn vị thí nghiệm. Điều kiện chăm sóc, nuôi dưỡng và quản lý vật nuôi (nếu thí nghiệm tiến hành trên vật nuôi). Số lần lặp lại là bao nhiêu? Tổ chức, quản lý và phân tích các số liệu thu thập được như thế nào? Mô hình phân tích số liệu ra sao?
- Kết quả dự kiến: Mặc dù chưa tiến hành nghiên cứu, tác giả cần dự kiến các kết quả đạt được. Lý tưởng nhất là các kết quả dự kiến được cụ thể hóa thành các bảng biểu và các biểu đồ nằm trong các mục khác nhau của phần kết quả và thảo luận. Trong một số trường hợp đặc biệt tác giả cần tạo ra các giá trị quan trắc giả định [có thể dùng hàm randbetween()] để biểu diễn kết quả dự kiến.
- Kế hoạch nghiên cứu: Đề cương nghiên cứu cần có kế hoạch nghiên cứu để giúp cho công tác giám sát và quản lý hoạt động một cách tốt nhất.

- Tiến hành nghiên cứu thu thập số liệu quan trắc

Sau khi đề cương được phê duyệt về mặt nội dung cũng như kinh phí nghiên cứu, tác giả tiến hành nghiên cứu để thu thập các số liệu quan trắc.

- Báo cáo kết quả nghiên cứu

Khi có các số liệu quan trắc, tác giả tiến hành xử lý số liệu và báo cáo kết quả nghiên cứu. Các số liệu thu thập được có thể xử lý theo một trong hai hướng hoặc kết hợp cả hai. Thống kê mô tả thông qua các tham số đặc trưng cho mức độ tập trung và

mức độ phân tán. Biểu đồ cũng là công cụ hữu hiệu để mô tả kết quả nghiên cứu. Tác giả nên linh động trong việc sử dụng bảng biểu và biểu đồ. Thống kê suy diễn được dùng để ước tính các chỉ số nghiên cứu cũng như diễn dịch các số liệu thu thập được.

Sau khi các số liệu đã được xử lý, tùy theo trường hợp mà tác giả có thể sử dụng các hình thức báo cáo sau:

- Bài báo khoa học đăng trên các tạp chí khoa học
- Kỷ yếu ở các hội nghị khoa học
- Bài trình bày tại các hội nghị khoa học
- Poster trong các hội nghị khoa học
- Báo cáo nghiệm thu đề tài
- Luận văn hoặc luận án tốt nghiệp

1.15. Bài tập

1. Nêu các biện pháp nâng cao độ chính xác của thí nghiệm?
2. Người nông dân A có 2 con bò bệnh, ông ta dùng thuốc B cho con bò thứ nhất, sau 1 tuần con bò đó hết bệnh, con bò thứ 2 không được dùng thuốc B thì phải hơn 1 tháng mới hết bệnh. Đây có phải là một thí nghiệm không? giải thích vì sao?
3. Người nông dân A có 12 con bò bị bệnh có triệu chứng giống nhau, trong đó có 6 con bò 2 năm tuổi và 6 con bò 3 năm tuổi. Người nông dân dùng thuốc B cho 3 con 2 năm tuổi và 3 con 3 năm tuổi, sau 1 tuần 6 con bò đó hết bệnh, 6 con bò còn lại không được dùng thuốc B thì sau 1 tháng mới hết bệnh. Hãy bình luận về thí nghiệm này?
4. Một nông dân A dùng một loại thuốc B cho một con bò bị bệnh của ông ta, sau 1 tuần con bò đó hết bệnh? Đây có phải là một thí nghiệm không? giải thích vì sao?
5. Người nông dân A có 6 con bò bị bệnh có triệu chứng giống nhau, người nông dân dùng thuốc B cho 3 con, sau 1 tuần 3 con này hết bệnh, 3 con còn lại không được dùng thuốc B thì sau 1 tháng mới hết bệnh. Hãy bình luận về thí nghiệm này?
6. Người nông dân A có 12 con bò bị bệnh có triệu chứng giống nhau, trong đó có 6 con bò 2 năm tuổi và 6 con bò 3 năm tuổi. Người nông dân dùng thuốc B cho 6 con 2 năm tuổi, sau 1 tuần 6 con bò đó hết bệnh, 6 con bò 3 năm tuổi còn lại không được dùng thuốc B thì sau 1 tháng mới hết bệnh. Hãy bình luận về thí nghiệm này?
7. Tại sao bất kỳ thí nghiệm nào đều yêu cầu tính ngẫu nhiên và tính lặp lại?
8. Tìm một bài báo trên tạp chí chuyên ngành (ví dụ: Nông nghiệp Phát triển Nông thôn), xác định giả thuyết nghiên cứu, nghiệm thức, đơn vị thí nghiệm, yếu tố khối (nếu có)?

9. Tìm một bài báo trên tạp chí chuyên ngành (ví dụ: Nông Nghiệp Phát Triển Nông Thôn) và bình luận là kết quả của nghiên cứu có xuất phát từ một nghiên cứu tốt hay không? Xem xét trên các khía cạnh
- a. Tổng quan tài liệu
 - b. Xác định vấn đề nghiên cứu
 - c. Giả thuyết nghiên cứu
 - d. Mục tiêu nghiên cứu
 - e. Thiết kế thí nghiệm: Thiết kế nghiên cứu đã phản ánh được giả thuyết nghiên cứu? mục tiêu nghiên cứu?
 - f. Đảm bảo yêu cầu của thí nghiệm: tính ngẫu nhiên, tính lặp lại, khống chế sự sai khác ban đầu (nếu có)?
 - g. Cách phân tích kết quả nghiên cứu có phù hợp không?
 - h. Kết luận có phù hợp không?
 - i. Định hướng nghiên cứu tiếp theo có không?
 - j. Danh mục tài liệu tham khảo có không?

Chương 2

THỐNG KÊ MÔ TẢ

Cốt lõi của chương này cung cấp các kiến thức liên quan đến thống kê mô tả. Cụ thể các tham số đặc trưng cho mức độ tập trung của mẫu như trung bình, trung vị và các tham số đặc trưng cho mức độ phân tán của mẫu như phương sai, độ lệch chuẩn. Phương pháp khác để đúc kết các số liệu quan trắc bằng cách sử dụng các biểu đồ như phân tán, hộp, cành và lá cũng được trình bày trong chương này.

2.1. Các tham số (thống kê) đặc trưng cho mức độ tập trung

2.1.1. Trung bình cộng

Trung bình cộng là tỷ số giữa tổng các giá trị quan trắc và số lượng các cá thể được quan trắc.

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} \quad [2.1]$$

Trong đó $\bar{X}$ là trung bình của mẫu có n quan trắc. Ở một số tài liệu khác, trung bình còn được ký hiệu là M. Trung bình quần thể được ký hiệu là μ . Trung bình là một thống kê tốt để đo lường mức độ tập trung mẫu khi sự phân phối của mẫu là đối xứng. Khi sự phân phối của mẫu là lệch thì trung bình không phải là thống kê tốt cho việc xác định mức độ tập trung, bởi vì giá trị trung bình bị ảnh hưởng lớn bởi các giá trị ngoại lệ. Ví dụ: các tính trạng sản xuất như tăng trọng (g/ngày), chiều cao vây (cm)... thường phân phối đối xứng; nhưng các tính trạng liên quan đến hàm lượng các chất, như hàm lượng phenol, hay methyl mercaptan trong chất thải từ chăn nuôi lợn thường có phân phối lệch. Trong trường hợp phân phối của tính trạng nghiên cứu là phân phối lệch, chúng ta nên dùng các loại tham số khác để đặc trưng cho mức độ tập trung, ví dụ như trung vị.

Trong trường hợp các số liệu phân nhóm, giá trị trung bình được tính theo công thức [2.2]

$$\bar{X} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} \quad [2.2]$$

Trong trường hợp mỗi mẫu nghiên cứu ứng với một giá trị trung bình, để tính giá trị trung bình chung, chúng ta có thể sử dụng công thức tính giá trị trung bình gia quyền [2.3].

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots n_k \bar{X}_k}{n_1 + n_2 + \dots n_k} \quad [2.3]$$

2.1.2. Trung bình nhân

Trung bình nhân còn gọi là trung bình hình học. Trung bình nhân chính là giá trị căn bậc n của tích n quan trắc.

$$\sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdot \dots \cdot X_n} = \text{anti log } \frac{1}{n} \sum_{i=1}^n \log X_i \quad [2.4]$$

Để hiểu hơn về giá trị trung bình nhân có thể xem xét ví dụ ở tập hợp số liệu sau:

Các bước tính trung bình nhân	X_i	$\ln(X_i)$
	1	0
	4	1,38629436
	2	0,69314718
	3	1,09861229
	80	4,38202663
Trung bình nhân từ số liệu thô	4,53586631	
Trung bình cộng từ số liệu ở dạng logarit tự nhiên ($e = 2,718$)		1,51201609
Exp(trung bình cộng từ số liệu ở dạng logarit tự nhiên)		4,53586631

Nếu sử dụng logarit cơ số 10 chúng ta cũng có kết quả trung bình nhân tương tự. Giá trị của trung bình nhân ít bị ảnh hưởng bởi các giá trị quan trắc ngoại lệ hơn so với trung bình cộng. Do vậy trung bình nhân thường được sử dụng để đo lường mức độ tập trung của các số liệu quan trắc có phân phối lệch.

Giá trị trung bình nhân thường được dùng để tính tỷ lệ bình quân. Ví dụ diễn biến tỷ lệ mắc bệnh dại trong thời gian từ năm 2005 đến 2009 như sau:

Năm	Tỷ lệ tăng so với năm trước (%)	Giá trị
2005	13	1,13
2006	22	1,22
2007	12	1,12
2008	- 5	0,95
2009	- 13	0,87

Giá trị trung bình nhân = 1,05. Điều này có nghĩa là tỷ lệ mắc bệnh đại bình quân hàng năm từ năm 2005 đến 2009 là 5%.

2.1.3. Trung vị

Trung vị là giá trị ở giữa của phân phối, một nửa các quan trắc trong phân phối có giá trị cao hơn trung vị và một nửa còn lại thấp hơn trung vị. Trung vị ít nhạy cảm với các giá trị ngoại lệ hơn so với tham số trung bình. Khi phân phối lệch tồn tại, để đặc trưng cho mức độ tập trung chúng ta nên sử dụng trung vị.

Khi phân phối là đối xứng thì tham số trung bình, trung vị và yếu số là giống nhau. Trung bình thường cao hơn trung vị trong phân phối lệch dương và thấp hơn trung vị trong phân phối lệch âm. Để xác định giá trị của tham số trung vị chúng ta chỉ cần sắp xếp các giá trị quan trắc theo thứ tự tăng dần hoặc giảm dần. Trong trường hợp tổng số các giá trị quan trắc là lẻ thì trung vị chính là giá trị ở giữa. Trong trường hợp tổng số các giá trị quan trắc là chẵn thì trung vị chính là trung bình của 2 giá trị ở giữa.

x_i	Sắp xếp tăng dần
3	2
4	3
5	3
7	3
3	4
3	Trung vị = 5
6	6
2	7
10	10
12	11
11	12

2.1.4. Yếu số

Yếu số là giá trị quan trắc có tần suất phân phối cao nhất. Sử dụng yếu số để đặc trưng cho mức độ tập trung. Yếu số có ưu điểm là không bị phụ thuộc vào các giá trị ngoại lệ do nó chỉ sử dụng một giá trị quan trắc có tần suất phân phối cao nhất. Yếu số còn có thể được sử dụng đối với các số liệu của biến định danh. Nhược điểm của yếu số là có thể có nhiều yếu số trong một phân phối.

Ví dụ: số lợn con cai sữa của 42 lợn nái ông bà dòng 1230 được thể hiện ở bảng 2.1. Hãy xác định yếu số.

Bảng 2.1. Số lợn con cai sữa (con/lứa) của lợn nái ông bà dòng C1230

Cá thể	Số con cai sữa	Cá thể	Số con cai sữa
1	8	22	8
2	12	23	9
3	7	24	8
4	11	25	10
5	9	26	10
6	14	27	7
7	6	28	9
8	12	29	9
9	12	30	10
10	11	31	8
11	9	32	10
12	11	33	9
13	8	34	12
14	11	35	7
15	9	36	8
16	9	37	9
17	8	38	10
18	8	39	10
19	9	40	12
20	10	41	8
21	10	42	9

Từ số liệu ở bảng 2.1, chúng ta có thể tính yếu số bằng cách tính tần suất phân phối (f).

Bảng 2.2. Tần suất phân phối của tính trạng số con cai sữa (con/lứa) của lợn ông bà dòng C1230

Số con cai sữa (con/lứa)	Tần suất (f)
6	1
7	3
8	9
9 = Yếu số	Tần suất cao nhất = 11
10	8
11	4
12	5
14	1

2.1.5. Trung bình lược giản

Trung bình lược giản (trimmed mean) được tính bằng cách loại bỏ một tỷ lệ nhất định các quan trắc cao nhất hoặc thấp nhất và tính trung bình cộng của phần còn lại. Ví dụ: trung bình lược giản 50% chính là trung bình của các quan trắc còn lại sau khi loại bỏ 25% quan trắc cao nhất và 25% quan trắc nhỏ nhất trong phân phối. Trung vị chính là trung bình lược giản 100% và trung bình cộng chính là trung bình lược giản 0%. Trung bình lược giản ít bị tác động bởi các quan trắc ngoại lệ hơn so với trung bình cộng. Trong các nghiên cứu đánh giá chất lượng thông qua cảm quan thì trung bình lược giản thường được sử dụng để hạn chế những đánh giá ngoại lệ.

2.1.6. Trung bình ba phần

Trung bình ba phần (trimean) được tính bằng cách cộng phân vị thứ 25% với 2 lần phân vị thứ 50% và với phân vị thứ 75% và chia cho 4.

$$\text{Trimean} = \frac{Q_1 + 2 \cdot Q_2 + Q_3}{4} \quad [2.5]$$

Trong đó Q_1 ; Q_2 và Q_3 lần lượt là phân vị thứ 25%; 50% và 75%. Ví dụ với tập hợp số liệu sau, chúng ta có $Q_1 = 51$; $Q_2 = 55$ và $Q_3 = 63$.

50	51	61
50	52	63
50	53	63
50	53	63
50	53	64
50	55	67
50	55	67
50	55	69
50	55	70
50	56	70
51	56	75
51	56	77
51	58	78
51	58	80
51	59	85
51	60	103
51	60	

$$\text{Trimean} = \frac{51 + 2 \times 55 + 63}{4} = 56$$

Trung bình ba phần đại diện cho mức độ tập trung của phân phối và có thể hạn chế tối đa của các quan trắc ngoại lệ, đặc biệt trong các phân phối lệch. Trong các trường hợp phân phối chuẩn, trung bình ba phần không nên sử dụng để đại diện cho mức độ tập trung của phân phối.

Tóm lại, để đánh giá mức độ tập trung của phân phối chúng ta có thể sử dụng một trong sáu tham số/thống kê trên. Mỗi tham số có ưu điểm và nhược điểm nhất định. Do vậy, tùy thuộc vào các phân phối cụ thể để lựa chọn tham số phù hợp nhất. Trung bình cộng thường được sử dụng nhiều nhất bởi vì trung bình sử dụng hết tất cả các quan trắc trong phân phối, do vậy về mặt lý thuyết trung bình cộng đại diện nhất cho mức độ tập trung của phân phối. Sử dụng trung bình cộng càng có hiệu quả khi phân phối là phân phối chuẩn. Tuy nhiên, khi phân phối là lệch thì sử dụng trung bình nhân có mức độ đại diện cho mức độ tập trung của phân phối hơn. Trung vị, yếu tố có thể thay thế trung bình để đại diện cho mức độ tập trung của phân phối trong trường hợp phân phối lệch. Tuy nhiên, nhược điểm của chúng là không sử dụng hết các quan trắc.

2.2. Các tham số (thống kê) đặc trưng cho mức độ phân tán

2.2.1. Khoảng biến thiên

Khoảng biến thiên là tham số đơn giản nhất đo lường sự phân tán của phân phối. Khoảng biến thiên chính là chênh lệch giữa giá trị cao nhất và giá trị thấp nhất của phân phối. Trong ví dụ về về số con cai sữa/lứa ở bảng 2.1, khoảng biến thiên = 14 - 6 = 8. Khoảng biến thiên ít được sử dụng để xác định mức độ phân tán do nó chỉ sử dụng hai giá trị quan trắc, do vậy bị ảnh hưởng lớn bởi các quan trắc ngoại lệ.

2.2.2. Khoảng biến thiên 50% ở giữa

Khoảng biến thiên 50% ở giữa (IQR) được sử dụng để đo lường mức độ biến thiên của phân phối. Giá trị của IQR được tính theo công thức [2.6].

$$IQR = Q_3 - Q_1 \quad [2.6]$$

IQR được dùng để xây dựng biểu đồ hộp. Trong trường hợp phân phối đối xứng IQR chính là trung vị.

2.2.3. Phương sai

Phương sai là thống kê phổ biến để đánh giá sự phân tán của phân phối. Nói cách khác, phương sai dùng để đánh giá mức độ biến thiên. Phương sai là cơ sở của một loạt thống kê suy diễn khác như phân tích phương sai. Phương sai mẫu được ký hiệu S^2 và σ^2 là phương sai quần thể. Công thức tính phương sai mẫu có dạng:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad [2.7]$$

Tử số của công thức [2.7] gọi là tổng các bình phương các độ lệch, gọi tắt là tổng các bình phương và được ký hiệu là SS. Trong trường hợp các giá trị quan trắc phân nhóm, chúng ta có thể dùng công thức [2.8] để tính phương sai:

$$S^2 = \frac{\sum_{i=1}^n f_i (X_i - \bar{X})^2}{n - 1} \quad [2.8]$$

Ví dụ: Hãy tính phương sai của tính trạng thời gian phối giống lại có kết quả sau cai sữa lợn con (ngày) của lợn nái ông bà dòng 1050, số liệu được thể hiện ở bảng 2.3.

Bảng 2.3. Thời gian phối giống lại có kết quả (ngày) của lợn nái sau cai sữa lợn con của lợn nái ông bà dòng C1050

Quan trắc thứ	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	57	32,56522	1060,493
2	32	7,565217	57,23251
3	31	6,565217	43,10208
4	42	17,56522	308,5369
5	26	1,565217	2,449905
6	6	- 18,4348	339,8412
7	5	- 19,4348	377,7108
8	7	- 17,4348	303,9716
9	85	60,56522	3668,146
10	5	- 19,4348	377,7108
11	5	- 19,4348	377,7108
12	4	- 20,4348	417,5803
13	113	88,56522	7843,798
14	6	- 18,4348	339,8412
15	8	- 16,4348	270,1021
16	5	- 19,4348	377,7108
17	4	- 20,4348	417,5803
18	47	22,56522	509,189
19	5	- 19,4348	377,7108
20	55	30,56522	934,2325
21	4	- 20,4348	417,5803
22	6	- 18,4348	339,8412
23	4	- 20,4348	417,5803
Trung bình	24,43478		
Tổng các bình phương			19579,65
Phương sai			889,9842

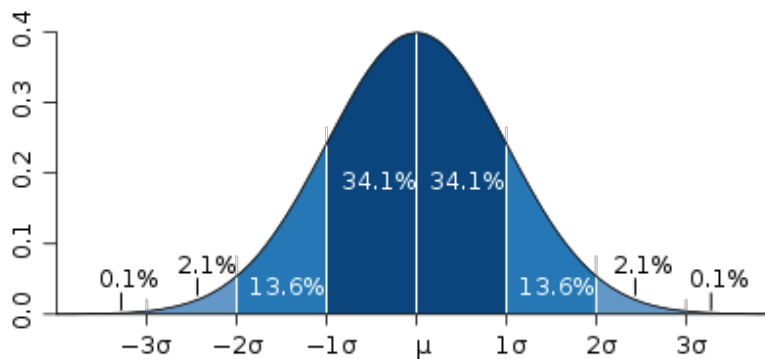
2.2.4. Độ lệch chuẩn

Độ lệch chuẩn là căn bậc hai của phương sai. Độ lệch chuẩn là tham số phổ biến nhất đo lường mức độ biến thiên và độ lệch chuẩn là cơ sở của một loạt các thống kê suy diễn. Độ lệch chuẩn của quần thể được ký hiệu là σ và độ lệch chuẩn của mẫu là s .

$$s^2 = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad [2.9]$$

Độ lệch chuẩn của tính trạng thời gian phối lại có kết quả sau cai sữa lợn con trong ví dụ ở bảng 2.3 là 29,8326. Chúng ta có thể đặt câu hỏi tại sao phải sử dụng độ lệch chuẩn, trong khi đó độ lệch chuẩn chẳng qua là căn bậc hai của phương sai. Sử dụng độ lệch chuẩn có tác dụng đưa đơn vị của tham số đo lường mức độ biến thiên của các quan trắc về cùng với đơn vị tính của tính trạng. Ví dụ ở bảng 2.3, phương sai = 889,9842, đơn vị của phương sai này không phải là ngày; trong khi đó độ lệch chuẩn của tính trạng này là 29,8326 có đơn vị là ngày.

Một ứng dụng rất quan trọng của độ lệch chuẩn trong việc xác định phân vị thứ của tính trạng nghiên cứu. Khi tính trạng nghiên cứu là phân phối chuẩn, chúng ta có thể nói 68% số quan trắc nằm trong khoảng một lần độ lệch tiêu chuẩn của trung bình, 95% số quan trắc nằm trong khoảng 2 lần độ lệch tiêu chuẩn của trung bình và 99,8% số quan trắc nằm trong khoảng 3 lần độ lệch chuẩn của trung bình (biểu đồ 2.1.). Cách thức xác định phân vị nêu trên gọi là quy tắc 3 σ . Từ quy tắc 3 σ chúng ta cũng có thể thấy rằng: $R \approx 6 \sigma$. Như vậy có thể ước tính độ lệch chuẩn từ R và ngược lại. Ước tính này càng chính xác khi mẫu quan trắc càng phân phối chuẩn.



Biểu đồ 2.1. Quy tắc 3 σ

Khi độ lệch chuẩn nhỏ thì các quan trắc phân phối gần với giá trị trung bình và khi độ lệch chuẩn lớn có nghĩa là các trị quan trắc nằm phân tán xa giá trị trung bình. Độ lệch chuẩn ít chịu sự tác động của các quan trắc ngoại lệ hơn so với khoảng biến thiên (R).

Một số tài liệu thống kê phân biệt 2 loại công thức tính phương sai và độ lệch chuẩn. Khi số lượng quan trắc < 30 thì mẫu số của công thức tính phương sai và độ lệch chuẩn lần lượt là $n - 1$ và $\sqrt{n-1}$; khi số số lượng quan trắc ≥ 30 thì mẫu số của công thức tính phương sai và độ lệch chuẩn lần lượt là n và $\sqrt{n}$. Khi số mẫu của công thức là n thì S^2 còn được gọi là ước tính chệch của σ^2 và khi mẫu số là $n - 1$ thì S^2 còn được gọi là ước tính không chệch của σ^2 .

Để đánh giá mức độ phân tán hay mức độ biến thiên của các số liệu quan trắc chúng ta có thể luôn luôn sử dụng mẫu số là $n - 1$ cho công thức tính phương sai. Lý do, mục đích của việc sử dụng phương sai là để đánh giá mức độ biến thiên của tính trạng nghiên cứu. Mức độ biến thiên này được tính dựa trên sự chênh lệch giữa mỗi giá trị quan trắc so với giá trị trung bình. Mỗi giá trị quan trắc đóng góp vào giá trị trung bình một lượng $\frac{1}{n}$. Để đánh giá một cách tinh khiết sự biến thiên của mỗi giá trị quan trắc so với phần còn lại của phân phối chúng ta cần loại bỏ sự đóng góp của giá trị quan trắc đó vào trong giá trị trung bình, lượng đó chính là $\frac{1}{n}x_i$. Do vậy mẫu số nên là $n - 1$ trong công thức tính phương sai và $\sqrt{n-1}$ trong công thức tính độ lệch chuẩn.

2.2.5. Hệ số biến dị

Hệ số biến dị (CV%) dùng để đo lường sự biến thiên của các tính trạng và so sánh sự biến thiên của các tính trạng có đơn vị đo lường khác nhau. Chúng ta không thể so sánh sự biến thiên của các tính trạng có đơn vị đo lường khác nhau thông qua tham số phương sai hay độ lệch chuẩn. Chúng ta có thể so sánh thông qua tham số tỷ lệ (không có đơn vị đo lường của tính trạng), ví dụ hệ số biến dị. Hệ số biến dị được tính qua công thức [2.10]. Trong ví dụ ở bảng 2.3, hệ số biến dị là 122,09%.

$$CV\% = \frac{S}{\bar{X}} \times 100 \quad [2.10]$$

Hệ số biến dị được dùng để đánh giá mức độ ổn định về năng suất, tính thuần chủng và khả năng thích nghi của các giống vật nuôi khác nhau.

2.2.6. Sai số của số trung bình

Sai số của số trung bình còn gọi là sai số chuẩn, ký hiệu là SE (Standard error of the mean) hay SEM. Sai số chuẩn được tính theo công thức [2.11].

$$SEM = \frac{S}{\sqrt{n}} \quad [2.11]$$

Trong đó: n là số lần lặp lại. Sai số chuẩn phản ánh mức độ đại diện của thống kê mẫu với tham số quần thể. Sai số của số trung bình được dùng để ước lượng trung

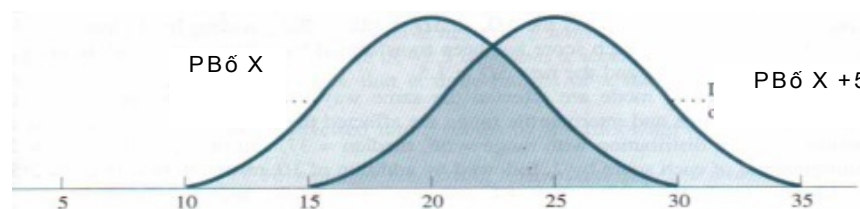
biên của quần thể, so sánh giữa hai giá trị trung bình và còn được dùng để xác định dung lượng mẫu. Sai số chuẩn được dùng để đánh giá mức độ biến thiên của giá trị trung bình.

Tại sao chúng ta cần tính SE. Nguyên lý và mục đích của thống kê học là ước tính những thông số của một quần thể. Trong thực tế chúng ta không biết các thông số này mà chỉ dựa vào những ước tính từ một hay nhiều mẫu để diễn dịch cho giá trị của quần thể từ các mẫu được chọn. Ví dụ để xác định các chỉ số vòng ngực và dài thân của lợn Móng Cái trưởng thành, chúng ta không thể xác định vòng ngực và dài thân của tất cả lợn Móng Cái trưởng thành ở Việt Nam, thay vào đó chúng ta phải chọn một mẫu gồm n cá thể để tính trị số trung bình của mẫu này và dùng trị số trung bình của mẫu để suy luận cho toàn quần thể. Mẫu càng ngẫu nhiên thì càng mang tính đại diện. Cứ mỗi lần chọn mẫu, chúng ta có một nhóm đối tượng khác. Và, cứ mỗi mẫu, chúng ta có một số trung bình mới. Nếu chúng ta chọn mẫu N lần (mỗi lần với n đối tượng) thì chúng ta sẽ có N số trung bình. Độ lệch chuẩn của N số trung bình này chính là sai số chuẩn. Do đó, sai số chuẩn phản ánh độ biến động của các số trung bình mẫu. Một số quan điểm cho rằng thuật ngữ “Standard error of the mean” = SEM là không đúng. Mà phải là Standard deviation of the means. Thay vì gọi Standard deviation of the means (quá dài dòng), người ta gọi ngắn gọn bằng một thuật ngữ mới Standard error.

Tóm lại, phương sai, độ lệch chuẩn, khoảng biến thiên, hệ số biến dị được dùng để đánh giá mức độ phân tán. Độ lệch chuẩn là tham số phổ biến nhất được dùng để đánh giá mức độ phân tán. Độ lệch chuẩn sử dụng hết các quan trắc trong mẫu và ít thay đổi do các quan trắc ngoại lệ hơn so với khoảng biến thiên.

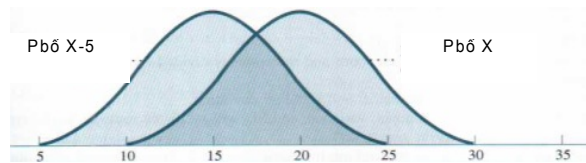
2.2.7. Sự chuyển dịch tuyến tính

Khi cộng thêm hoặc trừ một hằng số vào mỗi giá trị quan trắc của phân phối thì làm tăng hoặc giảm trung bình một lượng chính bằng hằng số, nhưng không làm thay đổi phương sai hoặc độ lệch chuẩn (biểu đồ 2.2; 2.3).

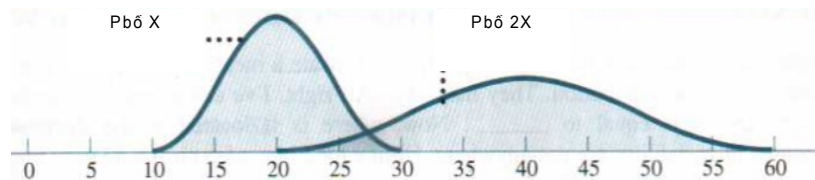


Biểu đồ 2.2: Sự chuyển dịch tuyến tính; cộng thêm một hằng số

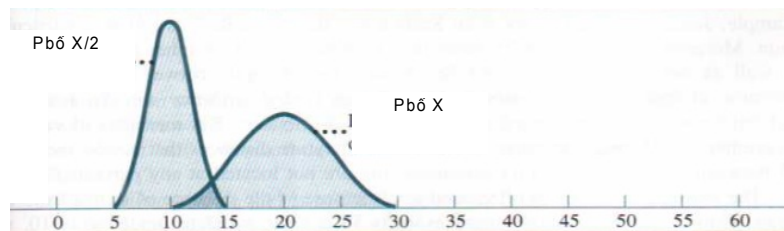
* Trừ đi một hằng số



* Nhân thêm một hằng số



* Chia cho một hằng số

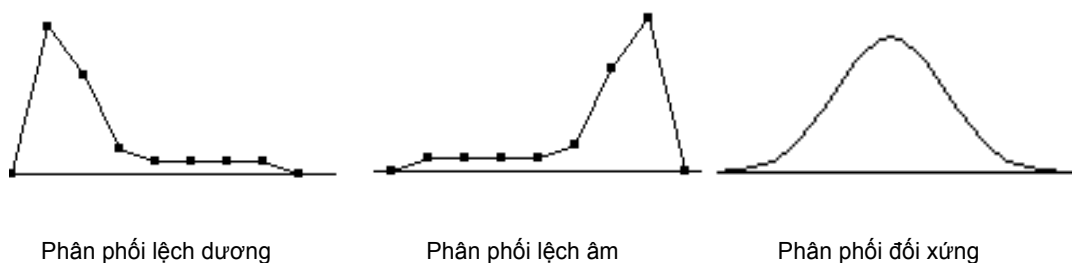


Biểu đồ 2.3: Sự chuyển dịch tuyến tính; trừ một hằng số, nhân thêm một hằng số và chia cho một hằng số

Khi nhân hay chia mỗi giá trị quan trắc một hằng số, trung bình và độ lệch chuẩn thay đổi một lượng tương tự như mỗi quan trắc thay đổi.

2.3. Hình dạng phân phối

Một phân phối được gọi là phân phối lệch khi một đuôi này dài hơn đuôi kia. Phân phối bên trái trong biểu đồ 2.4 là phân phối lệch dương, có nghĩa là đuôi của phân phối kéo dài về hướng dương tính. Phân phối ở giữa là phân phối lệch âm do đuôi của phân phối kéo dài về phía âm tính. Phân phối bên phải là đối xứng.



Biểu đồ 2.4. Các hình dạng phân phối

Để xác định độ lệch của phân phối chúng ta tính giá trị skewness. Khi phân phối là chuẩn thì giá trị skewness = 0. Khi trung bình lớn hơn trung vị thì phân phối lệch dương và ngược lại. Để tính độ lệch của phân phối chúng ta có thể sử dụng công thức 2.12.

$$\text{Skew} = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{N \times S^3} \quad [2.12]$$

Ngoài ra để phản ánh độ nhọn của phân phối chúng ta có thể tính giá trị kurtosis. Khi phân phối chuẩn thì giá trị kurtosis của phân phối = 0.

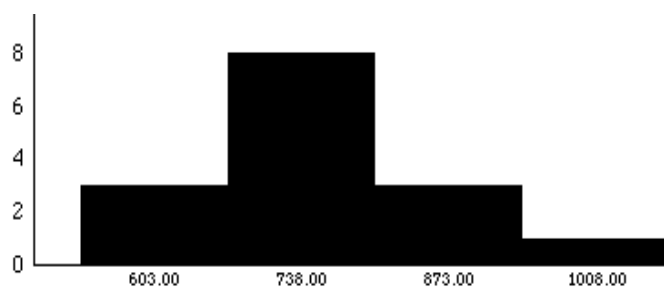
$$\text{Kurtosis} = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{N \times S^4} \quad [2.13]$$

2.4. Trình bày các quan trắc bằng biểu đồ

Trình bày các quan trắc bằng các biểu đồ có ý nghĩa lớn trong việc phát hiện khuynh hướng, từ đó định hướng xử lý số liệu tiếp theo. Phần này giới thiệu một số loại biểu đồ phổ biến dùng để đúc kết các số liệu quan trắc như tổ chức đồ, biểu đồ phân tán, biểu đồ hộp...

2.4.1. Tổ chức đồ

Tổ chức đồ được xây dựng từ bảng tần suất. Khoảng cách được thể hiện trên trục hoành và số lượng các quan trắc trong mỗi khoảng cách được thể hiện bằng chiều cao của hình chữ nhật phía trên khoảng cách. Ví dụ: tăng trọng (g/ngày) của 15 lợn thịt lai 3 máu (Omega × Yorkshire) trong thời gian nuôi thịt từ cai sữa đến xuất chuồng 90 kg như sau: 700; 700; 700; 600; 717; 550; 867; 1067; 883; 567; 833; 683; 800; 500; 683. Từ các giá trị tăng trọng này chúng ta có thể xây dựng tổ chức đồ như sau:

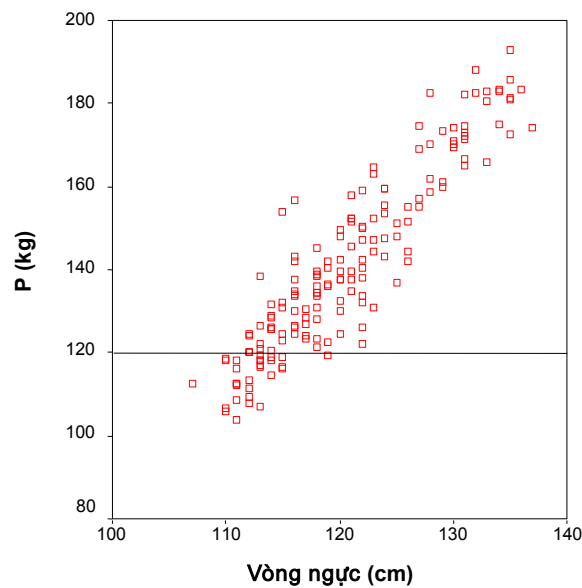


Biểu đồ 2.5. Tổ chức đồ thể hiện tăng trọng của lợn thịt 3 máu (Omega × Yorkshire)

Hình dạng của tổ chức đồ thay đổi rất lớn phụ thuộc vào sự lựa chọn kích cỡ của khoảng cách trên trục hoành. Biểu đồ dạng thanh cũng có hình dạng tương tự, chỉ khác là các thanh được cách nhau ra và đồ thị dạng thanh thường được sử dụng cho các biến định tính. Tổ chức đồ có thể được dùng để suy đoán phân phối chuẩn.

2.4.2. Biểu đồ phân tán

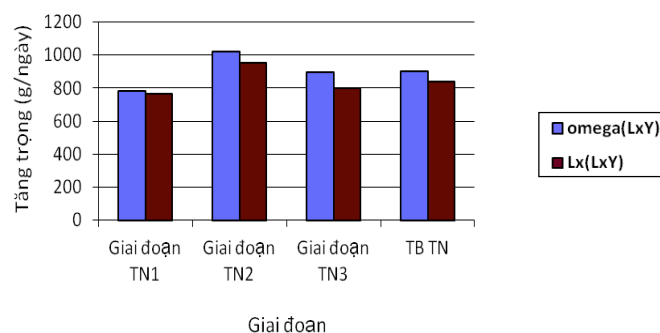
Biểu đồ phân tán được sử dụng để phát hiện mối quan hệ giữa 2 biến. Biểu đồ phân tán được dùng để phát hiện quy luật phân phối của trình trạng trước khi xác định phương pháp xử lý số liệu thích hợp. Ví dụ từ biểu đồ 2.6 chúng ta có thể suy đoán mối quan hệ tuyến tính giữa vòng ngực (cm) và khối lượng của bò trong khoảng vòng ngực và khối lượng xác định. Biểu đồ phân tán còn được dùng để kiểm tra các giá trị ngoại lệ, đặc biệt là khi số lượng các quan trắc lớn. Mỗi điểm trên biểu đồ phân tán là sự kết hợp giữa vòng ngực và khối lượng xác định trên cũng một cá thể.



Biểu đồ 2.6. Biểu đồ phân tán biểu diễn mối quan hệ giữa vòng ngực (cm) và khối lượng của bò (kg)

2.4.3. Biểu đồ dạng thanh

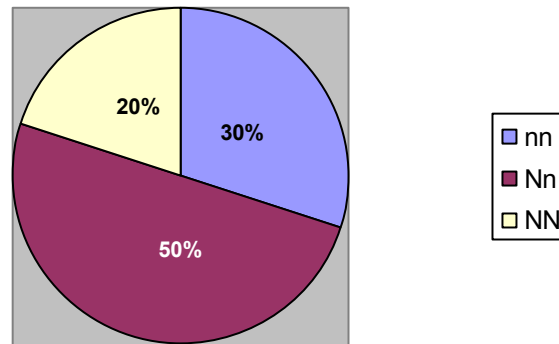
Đồ thị dạng thanh thường được sử dụng cho các biến độc lập là biến định tính.



Biểu đồ 2.7. Biểu đồ dạng thanh so sánh tăng trọng của lợn thịt lai Omega × (Landrace × Yorkshire) và Landrace × (Landrace × Yorkshire)

2.4.4. Biểu đồ dạng bánh

Biểu đồ dạng bánh thường được dùng cho các số liệu tỷ lệ. Ví dụ khảo sát 1000 lợn Pietrain sơ sinh để phát hiện 3 kiểu gen halothan, kết quả được thể hiện ở biểu đồ 2.8.



Biểu đồ 2.8. Biểu đồ dạng bánh về tần số kiểu gen Halothan của lợn sơ sinh Pietrain

2.4.5. Biểu đồ dạng cành và lá

Biểu đồ dạng cành và lá là một trong những phương pháp biểu diễn các quan trắc bằng hình ảnh. Biểu đồ này có hiệu quả khi các số liệu quan trắc quá lớn. Biểu đồ cành và lá gần giống như tổ chức đồ. Ví dụ tập hợp số liệu 85,3; 80,3; 75,9; 74,9; 71,1; 60; 58,1; 56,4; 51,2; 45,6 và 40,1 có thể có biểu đồ cành và lá như sau:

Cành	Lá
8	05
7	156
6	0
5	168
4	06

Nhân 10 với cành

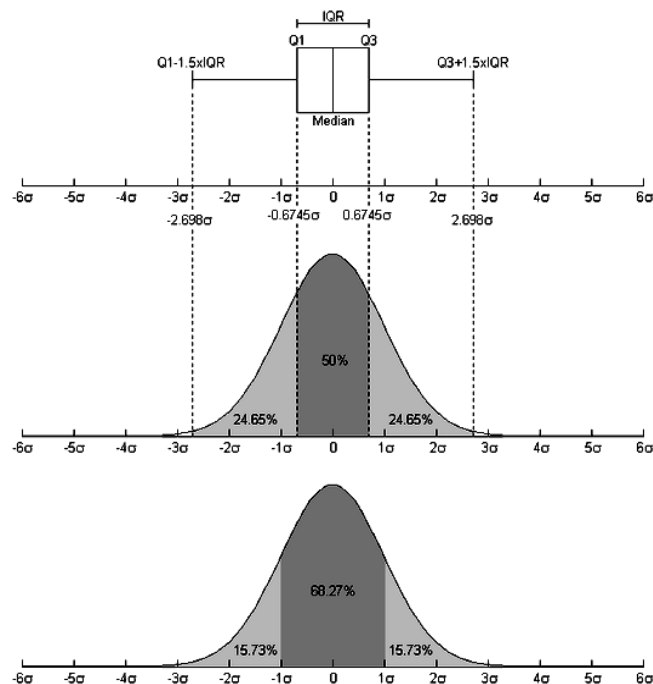
Biểu đồ 2.9. Biểu đồ cành lá

Giá trị quan trắc 85,3 có thể xác định gần đúng như sau: $10 \times 8 + 5$. Trên biểu đồ phần cành thể hiện 8 và phần lá thể hiện 5. Số 5 được thể hiện ở cột đầu tiên từ bên phải sang. Khi cành chỉ có một giá trị thì lá có thể nhận giá trị từ 0 đến 9 (biểu đồ 2.9).

2.4.6. Đồ thị dạng hộp

Đồ thị dạng hộp là một công cụ hình ảnh tuyệt vời để tóm tắt các khía cạnh khác nhau của phân phối. Hộp kéo dài từ phân vị thứ 25% đến phân vị thứ 75%, điều này có nghĩa hộp chứa 50% quan trắc ở giữa; 25% quan trắc nằm phía dưới hộp và 25% quan trắc nằm phía trên hộp. Giá trị trung vị chính là đường chạy qua hộp, nếu đường ngang này không nằm giữa hộp thì sự phân phối không đối xứng mà là phân phối lệch. Điểm cuối của 2 râu thể hiện giá trị tối đa và tối thiểu của phân phối, trong trường hợp không có các quan trắc ngoại lệ.

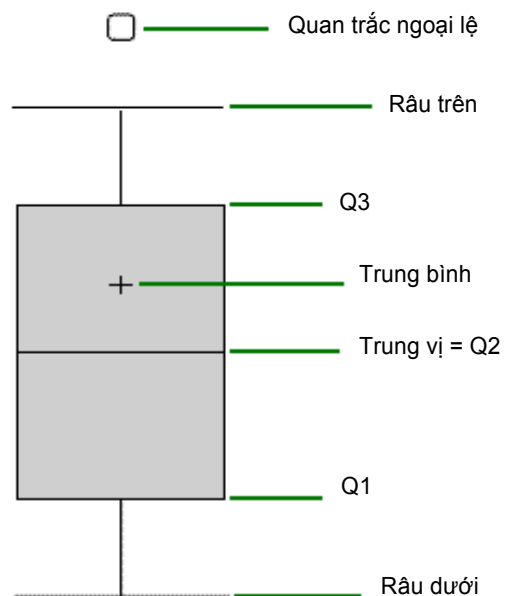
Trong trường hợp phân phối có các quan trắc ngoại lệ thì 2 râu được kéo dài đến $Q3 + 1,5 \times IQR$ và $Q1 - 1,5 \times IQR$. Các quan trắc định vị bên ngoài râu thì nên xem xét về tính ngoại lệ. Thông thường các quan trắc ngoại lệ này được thể hiện là chấm đen “●”, hoặc hình ngôi sao hoặc là vòng tròn nhỏ “o”. Một vài biểu đồ hộp có thể có một dấu chéo (x) hoặc một chấm tròn bên trong hộp (từ $Q1$ đến $Q3$) để thể hiện giá trị trung bình.



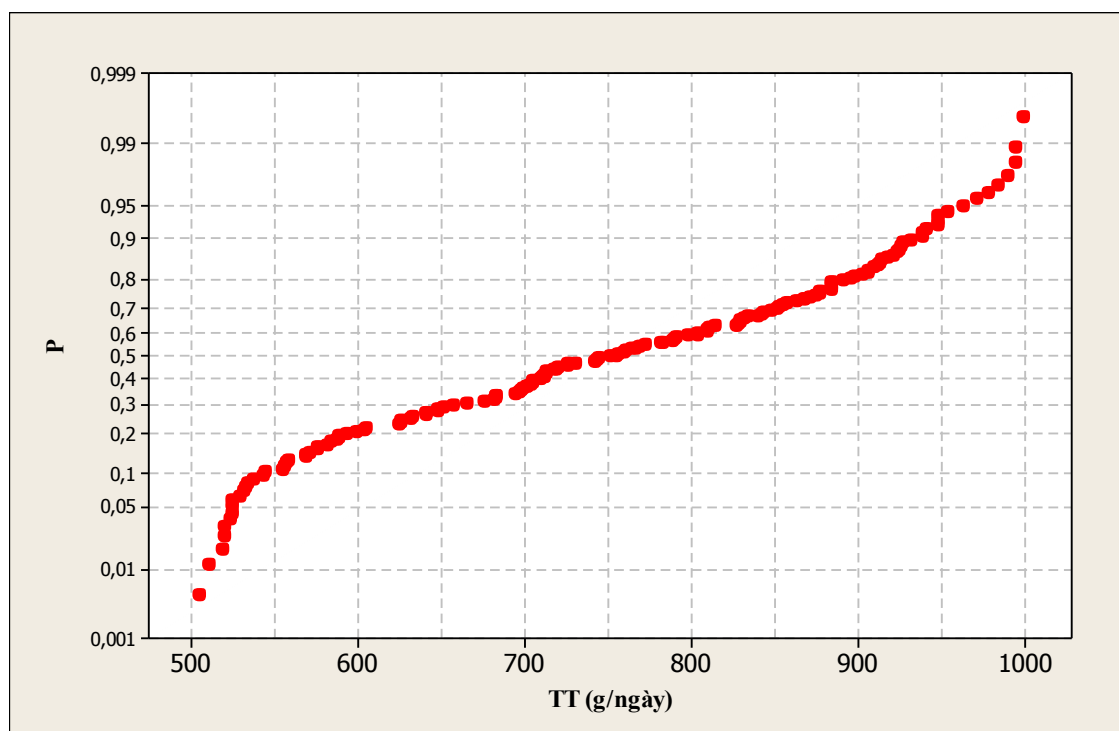
Biểu đồ 2.10. Mối quan hệ giữa Q, IQR và σ

2.4.7. Biểu đồ xác suất tích lũy

Biểu đồ xác suất tích lũy dùng để xác định xác suất tích lũy của tính trạng nghiên cứu. Giá trị trên trục hoành là giá trị của tính trạng nghiên cứu và giá trị trên trục tung là xác suất tích lũy. Tổng xác suất tích lũy tất nhiên bằng 1. Qua biểu đồ 2.12 ta thấy rằng có khoảng 80% cá thể lợn có tăng trọng (g/ngày) nhỏ hơn 900 g/ngày.



Biểu đồ 2.11. Biểu đồ mô tả các thành phần trong biểu đồ hộp



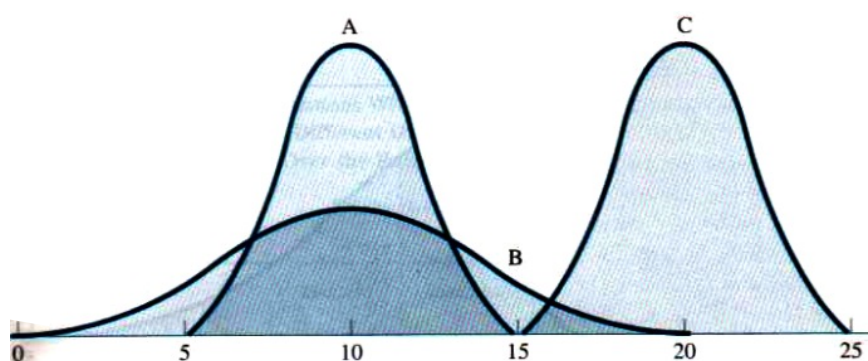
Biểu đồ 2.12. Biểu đồ xác suất tích lũy của tăng trọng của lợn lai $\frac{3}{4}$ máu ngoại Pietrain × (Yorkshire × Móng Cái) trong thời gian nuôi thịt từ 20 đến 100 kg

2.5. Bài tập

1. Hãy tính các tham số đặc trưng cho mức độ tập trung, mức độ phân tán và trình bày bằng các loại biểu đồ của tập hợp số liệu tăng trọng (g/ngày) của lợn lai 3/4 máu ngoại Pietrain $\times$ (Yorkshire $\times$ Móng Cái) trong thời gian nuôi thịt từ 20 đến 100 kg

604	913	885	885	635	597	639	948
524	1000	982	863	729	594	803	666
644	518	809	961	888	652	959	715
807	893	516	620	952	807	918	898
942	503	897	751	680	702	667	627
849	522	789	721	835	989	841	849
892	563	519	994	574	854	775	586
859	541	866	533	576	749	744	570
747	672	854	664	760	514	821	688
558	664	677	549	566	949	866	891
571	622	767	643	740	835	563	728
574	600	919	841	522	937	833	712
894	864	834	563	964	697	572	988
910	811	985	824	862	699	617	519
896	982	542	885	528	623	969	638

Hãy so sánh 2 phân bố A với B; A với C?



2. Tại sao để mô tả một biến số sinh học theo luật phân phối chuẩn, các nhà nghiên cứu thường dùng cách trình bày số trung bình và kèm độ lệch chuẩn mà không phải là sai số chuẩn?

3. Khi độ lệch chuẩn lớn hơn giá trị trung bình, chúng ta có thể có suy đoán gì về phân phối của tính trạng nghiên cứu?

4. Cho 3 tập hợp số liệu sau

- Tập hợp số liệu 1: 8 8 10 11 13
- Tập hợp số liệu 2: 8 8 10 11 200
- Tập hợp số liệu 3: - 100 8 10 11 13

Đề đặc trưng cho mức độ tập trung, trong mỗi tập hợp số liệu nên sử dụng loại tham số nào?

5. Hãy nêu mối quan hệ giữa độ lệch tiêu chuẩn và khoảng biến động? Khi nào thì ước tính độ lệch chuẩn từ khoảng biến động có độ chính xác cao?

6. Tại sao khi mô tả một biến số sinh học chúng ta phải sử dụng cả tham số đặc trưng cho mức độ tập trung và tham số đặc trưng cho mức độ phân tán?

Chương 3

KIỂM TRA GIẢ THUYẾT VÀ LỰC THỐNG KÊ

Khi thiết kế thí nghiệm và phân tích số liệu, chúng ta thường đặt các câu hỏi như thí nghiệm cần bao nhiêu lần lặp lại? chúng ta cũng có thể muốn biết lực thống kê của kiểm tra để phát hiện một ảnh hưởng có ý nghĩa với dung lượng mẫu và mức độ tin cậy cho trước? chúng ta cũng có thể có câu hỏi về dung lượng mẫu tối thiểu là bao nhiêu với lực thống kê và mức độ tin cậy cho trước. Để trả lời các câu hỏi đó, chương này nhằm mục tiêu cung cấp kiến thức liên quan đến giả thuyết, kiểm tra giả thuyết, sai sót loại I, sai sót loại II, lực thống kê và số lần lặp lại cần thiết.

3.1. Kiểm tra giả thuyết

3.1.1. Khái niệm và phân loại giả thuyết

Giả thuyết là một nhận định cụ thể về một đặc tính của quần thể nghiên cứu. Giả thuyết cũng có thể xem là một kết luận sơ bộ về một đặc tính của quần thể nghiên cứu. Kiểm tra giả thuyết là một tiến trình khoa học kiểm tra một giả thuyết có hợp lý hay không.

Ví dụ về một giả thuyết “*giảm mức protein trong khẩu phần của lợn nuôi thịt từ 17% xuống 13% và bổ sung các axit amin thiết yếu dạng tinh thể không làm giảm tăng trọng của lợn*”. Kiểm tra giả thuyết là tiến trình chứng minh giả thuyết này đúng hay sai. Tiến trình đó bao gồm nhiều công đoạn khác nhau như tiến hành thí nghiệm trên lợn, thu thập, xử lý số liệu....

Khi nêu giả thuyết cần chú ý một số điểm sau:

- Một giả thuyết nên hấp dẫn và gây sự chú ý cho độc giả và quan trọng là giả thuyết phải có thể kiểm tra được. Không nên có những giả thuyết không thể hoặc chưa thể kiểm tra được, ví dụ *con lai giữa bò và cừu có tăng trọng cao hơn trung bình bố và mẹ nó*. Trong điều kiện hiện tại chúng ta chưa có thể tạo ra tổ hợp lai này.
- Không nên có những giả thuyết mang tính hiển nhiên, ví dụ *nước được cấu tạo từ hydro và oxy*, đây là kiến thức đã được chứng minh từ lâu và đã trở thành hiển nhiên.
- Một giả thuyết là một nhận định/ước đoán/phỏng đoán về một tham số của quần thể, chứ không phải là của một mẫu. Chúng ta không thể có giả thuyết $H_0: \bar{x} = 0$ bởi vì chúng ta có thể tính được trung bình của mẫu từ mẫu nghiên cứu.
- Cần phân biệt giữa giả thuyết và giả định. Giả định là điều kiện để đạt được một mục tiêu nào đó. Ví dụ lợn lai F_1 (Pietrain $\times$ Yorkshire) sẽ đạt tăng trọng khoảng 750 g/ngày nếu được nuôi trong điều kiện chăn nuôi công nghiệp.

- Giả thuyết gồm có hai loại H_0 và H_A hay H_1 . Giả thuyết H_0 là nhận định cụ thể về một đặc tính của quần thể nghiên cứu và thường có dạng *không có ảnh hưởng*, hoặc *không sai khác giữa hai hoặc nhiều hơn hai quần thể nghiên cứu*. Theo Murphy (1998) lý do giả thuyết H_0 có dạng như vậy là để dễ tính toán kiểm tra thống kê (standardized effect sizes) và diễn dịch kết quả kiểm tra. Giả thuyết H_A hay H_1 còn gọi là giả thuyết thay thế, là sự phản bác giả thuyết H_0 hay còn gọi là đối thuyết. Các nhà nghiên cứu thường nhưng không phải lúc nào cũng mong đợi các bằng chứng để chấp nhận đối thuyết.
- Giả thuyết có thể là hai đuôi ($H_0: \mu = 0$) hoặc một đuôi ($H_0: \mu \geq 0$ hoặc $H_0: \mu \leq 0$). Giả thuyết một đuôi hoặc hai đuôi liên quan đến kiểm tra một đuôi hoặc hai đuôi. Kiểm tra một đuôi xem xét một trong hai bên của sự phân bố. Kiểm tra hai đuôi xem xét cả hai đuôi của sự phân bố.

Nhìn chung, kiểm tra giả thuyết có thể có 5 bước sau:

Bước 1: Nêu giả thuyết H_0 , chú ý nêu rõ là giả thuyết một đuôi hay hai đuôi.

Bước 2: Lựa chọn mức độ tin cậy (mức ý nghĩa). Chú ý là giả thuyết một đuôi hay hai đuôi để từ đó xác định chính xác giá trị tra bảng và vùng phản bác/vùng bác bỏ

Bước 3: Tính toán các giá trị kiểm tra thống kê và giá trị P hay xác định khoảng tin cậy

Bước 4: Quyết định bác bỏ hay không bác bỏ giả thuyết H_0 bằng cách so sánh giá trị ở bước 2 và bước 3 (giá trị tra bảng và giá trị tính toán)

Bước 5: Kết luận và diễn dịch kết quả

3.1.2. Phương pháp tiếp cận trong kiểm tra giả thuyết

Có ba phương pháp tiếp cận trong kiểm tra giả thuyết. Cả ba phương pháp tiếp cận đều cho cùng một kết quả (Bảng 3.1).

Bảng 3.1. Phương pháp tiếp cận trong kiểm tra giả thuyết

Bước	Phương pháp tiếp cận giá trị kiểm tra thống kê	Phương pháp tiếp cận giá trị xác suất	Phương pháp tiếp cận khoảng tin cậy
1	Nêu giả thuyết H_0 và H_A	Nêu giả thuyết H_0 và H_A	Nêu giả thuyết H_0 và H_A
2	Xác định giá trị mức độ tin cậy α và tra bảng xác định giá trị thống kê tra bảng hay còn gọi giá trị tới hạn = CV (ví dụ giá trị t, F, χ^2)	Xác định giá trị mức độ tin cậy α	Xác định giá trị mức độ tin cậy α hoặc $1 - \alpha$ và một giá trị giả thuyết
3	Tính toán giá trị kiểm tra thống kê (giá trị tính toán = TS)	Tính toán giá trị kiểm tra và giá trị xác suất P của nó	Xác định khoảng tin cậy $(1 - \alpha)100\% = CI$
4	Nếu giá trị TS > giá trị CV, bác bỏ giả thuyết H_0	Nếu $P < \alpha$, bác bỏ giả thuyết H_0	Nếu giá trị giả thuyết không nằm trong CI, bác bỏ giả thuyết H_0
5	Kết luận và diễn dịch	Kết luận và diễn dịch	Kết luận và diễn dịch

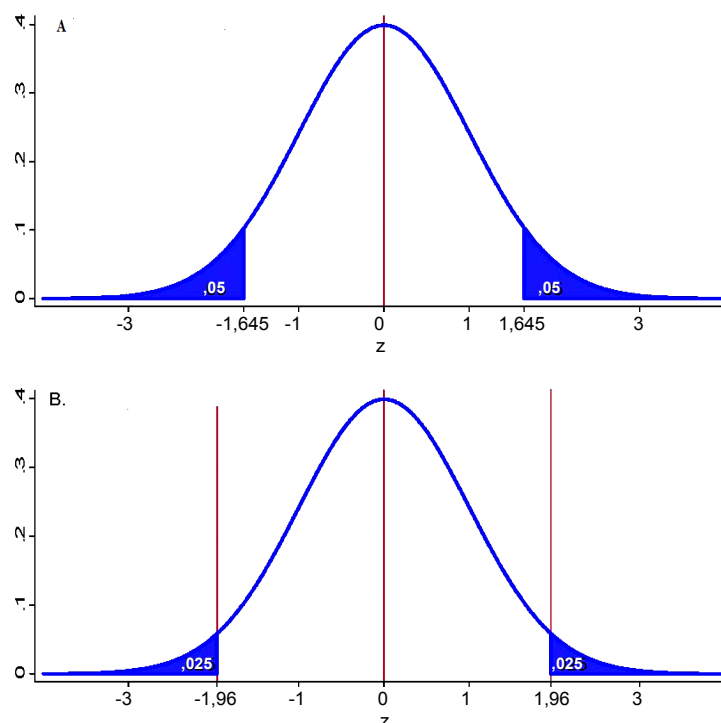
Trong phương pháp tiếp cận giá trị kiểm tra thống kê, chúng ta tính toán giá trị kiểm tra từ các số liệu thí nghiệm còn gọi là giá trị tính toán (TS) và so sánh giá trị đó với giá trị tra bảng hay còn gọi là giá trị quyết định (CV) hoặc giá trị tới hạn. Nếu giá trị tính toán lớn hơn giá trị tra bảng hoặc nếu giá trị tính toán nằm trong vùng phản bác thì giả thuyết H_0 bị bác bỏ. Trong phương pháp tiếp cận xác suất, chúng ta tra bảng giá trị xác suất P từ giá trị TS và so sánh với mức độ tin cậy α . Nếu giá trị P nhỏ hơn α , chúng ta bác bỏ giả thuyết H_0 . Trong phương pháp tiếp cận khoảng tin cậy, nếu giá trị giả thuyết không nằm trong khoảng tin cậy thì chúng ta bác bỏ giả thuyết H_0 .

3.1.3. Sai sót loại I và mức ý nghĩa

Mức ý nghĩa là xác suất gặp phải sai sót loại I. Sai sót loại I xảy ra khi H_0 bị bác bỏ mặc dù H_0 là đúng (Bảng 3.2). Mức ý nghĩa được ký hiệu là α và $1 - \alpha$ là khoảng tin cậy.

Bảng 3.2. Sai sót loại I, loại II và lực thống kê

Số liệu	Sự thực quần thể	
	H_0 là đúng	H_0 là sai
Chấp nhận H_0	$1 - \alpha = 95\%$	Sai sót kiểu II = $\beta = 20\%$
Loại bỏ H_0	Sai sót kiểu I = $\alpha = 5\%$	Lực thống kê = $1 - \beta = 80\%$

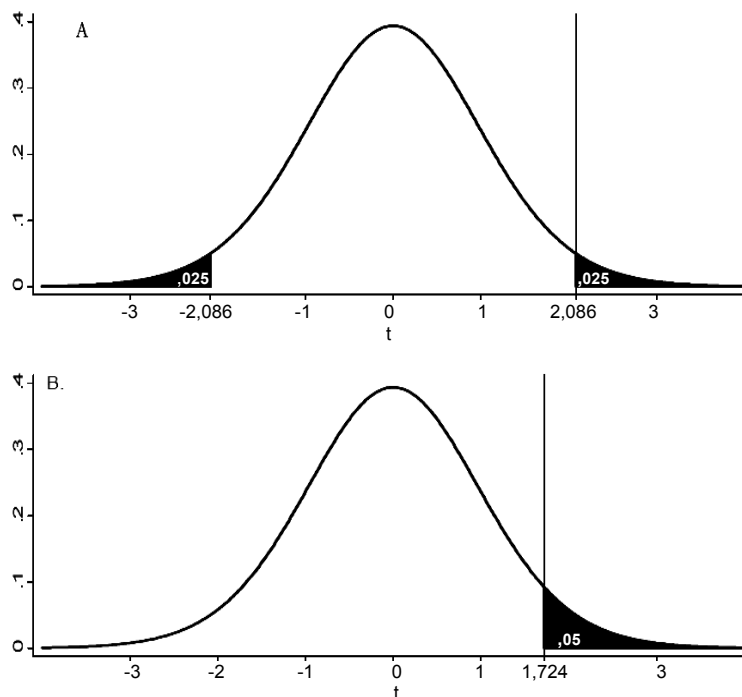


Biểu đồ 3.1. Mức ý nghĩa, giá trị quyết định (CV) trong phân bố chuẩn

Trong một kiểm tra hai đuôi, mức ý nghĩa là tổng của hai phần diện tích đối xứng tại hai đuôi của sự phân bố. Biểu đồ 3.1 trình bày ví dụ về vùng bác bỏ. Các vùng được bao bọc quanh bởi đường phân bố xác suất, trục hoành và giá trị quyết định (CV) thì được gọi là vùng bác bỏ. Điều này có nghĩa là giả thuyết H_0 sẽ bị bác bỏ nếu giá trị tính toán (TS) nằm trong các vùng bác bỏ này.

Trong hình phía trên của biểu đồ 3.1, chúng ta có mức ý nghĩa là 0,1. Giá trị 0,1 này là tổng của 2 vùng bác bỏ. Diện tích phía dưới đường phân bố xác suất từ một giá trị cụ thể, ví dụ 1,645 đến vô cùng dương là 0,05. Và tương tự diện tích phía dưới đường phân bố xác suất từ -1,645 đến âm vô cùng là 0,05. Giá trị cụ thể -1,645 hay 1,645 còn được gọi là giá trị quyết định (CV) hay giá trị tới hạn của mức ý nghĩa. Trong hình phía dưới của biểu đồ 3.1, giá trị -1,96 và 1,96 là hai giá trị tới hạn của mức ý nghĩa 0,05 và hai giá trị này hình thành nên hai vùng bác bỏ từ mỗi giá trị đến âm vô cùng và dương vô cùng và tổng diện tích phía dưới đường phân bố xác suất của cả hai vùng là 0,05. Khi mức ý nghĩa càng nhỏ, giá trị tới hạn CV càng dịch chuyển về hai cực của sự phân bố và vùng bác bỏ càng nhỏ dần và cơ hội bác bỏ giả thuyết H_0 càng thấp.

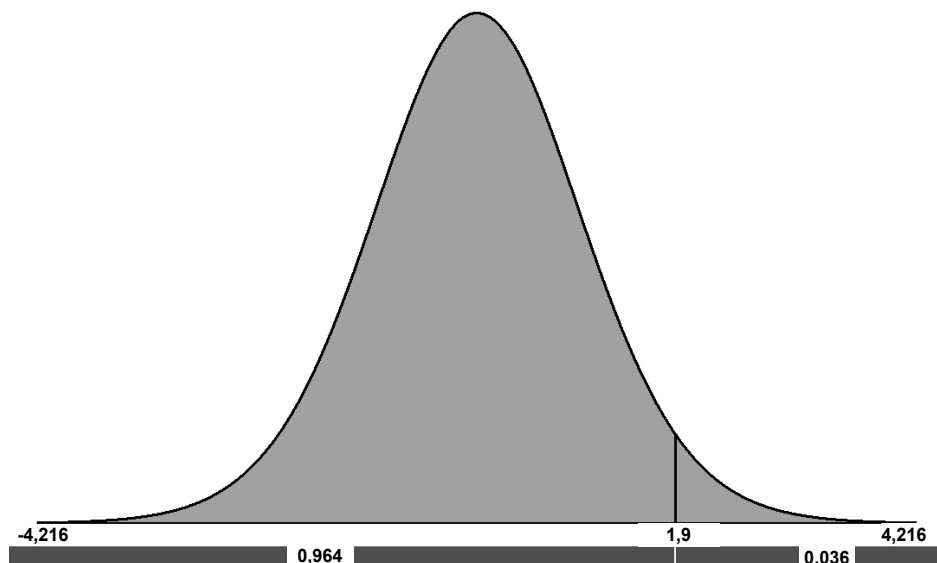
Giá trị tới hạn CV phụ thuộc vào phân bố xác suất, độ tự do, kiểm tra một đuôi hay kiểm tra hai đuôi. Ví dụ giá trị tới hạn trong trường hợp kiểm tra hai đuôi với độ tự do 20 là -2,068 và 2,068 ở mức ý nghĩa 0,05. Khi xem xét kiểm tra một đuôi về phía dương tính, giá trị CV là 1,724 ở mức ý nghĩa 0,05. Điều này dễ hiểu bởi vì diện tích phía dưới đường phân bố từ giá trị CV 1,724 đến 2,068 là 0,025, có nghĩa là diện tích từ CV = 1,724 đến vô cùng dương là $0,05 = 0,025 + 0,025$ (Biểu đồ 3.2).



Biểu đồ 3.2. Mức ý nghĩa, giá trị quyết định (CV) trong phân bố t

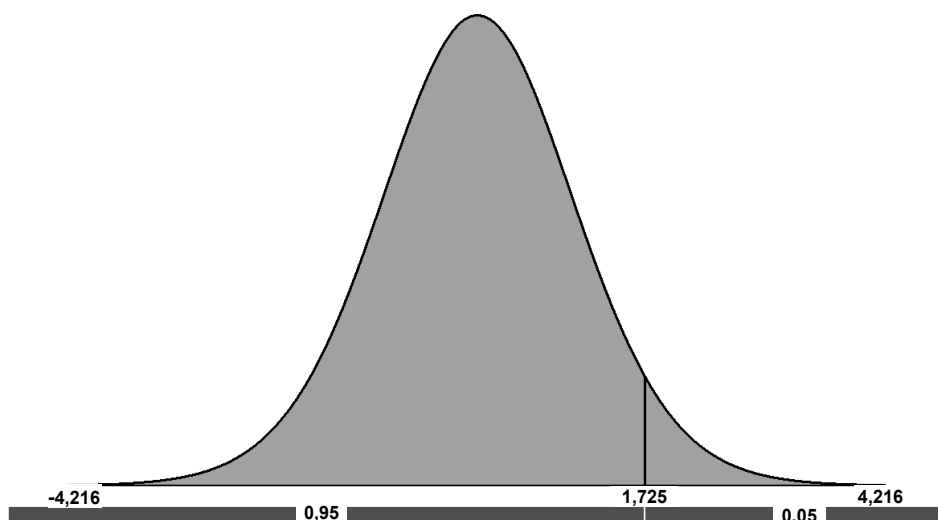
Mức ý nghĩa là mức độ mà chúng ta sẵn sàng chấp nhận rủi ro đưa ra quyết định sai. Mức 0,05 có nghĩa là chúng ta chấp nhận một rủi ro quyết định sai 5 lần trong 100 lần thử. Một cách hiểu khác của mức ý nghĩa $\alpha = 0,05$ có nghĩa là sự sai khác giữa hai giá trị trung bình có đến 5% là do yếu tố ngẫu nhiên, 95% là do yếu tố thí nghiệm. Khi chúng ta tăng giá trị mức ý nghĩa ví dụ từ 0,05 lên 0,1 thì chúng ta dễ dàng bác bỏ giả thuyết H_0 hơn, tuy nhiên kết luận của chúng ta ít thuyết phục hơn. Ngược lại, khi chúng ta giảm mức ý nghĩa từ 0,1 xuống 0,05 thì chúng ta khó bác bỏ giả thuyết H_0 hơn, tuy nhiên bù lại kết luận của chúng ta có tính thuyết phục hơn, ít rủi ro hơn.

Xem xét phân bố thứ nhất trong biểu đồ 3.2, ví dụ với giá trị tính toán (TS) là 1,9 thì chúng ta sẽ không bác bỏ giả thuyết H_0 ở mức ý nghĩa 0,05 (Biểu đồ 3.3), bởi vì 7,2% (Giá trị P) = diện tích từ hai giá trị TS đi về hai cực của sự phân bố (kiểm tra hai đuôi). Giả sử mức ý nghĩa tăng lên từ 0,05 đến 0,1 (có nghĩa chúng ta sẵn sàng chấp nhận rủi ro lên đến 10% đưa ra quyết định sai) thì chúng ta bác bỏ giả thuyết H_0 , bởi vì chúng ta sẵn sàng chấp nhận rủi ro đưa ra quyết định sai lên đến 10% trong khi đó giá trị tính toán TS chỉ cho phép lên đến 7,2%. Các số liệu nghiên cứu đã chỉ ra rằng ít rủi ro hơn (7,2%) so với dự kiến 10% để bác bỏ giả thuyết H_0 .



Biểu đồ 3.3. Giá trị P với giá trị tính toán 1,9 và độ tự do 20

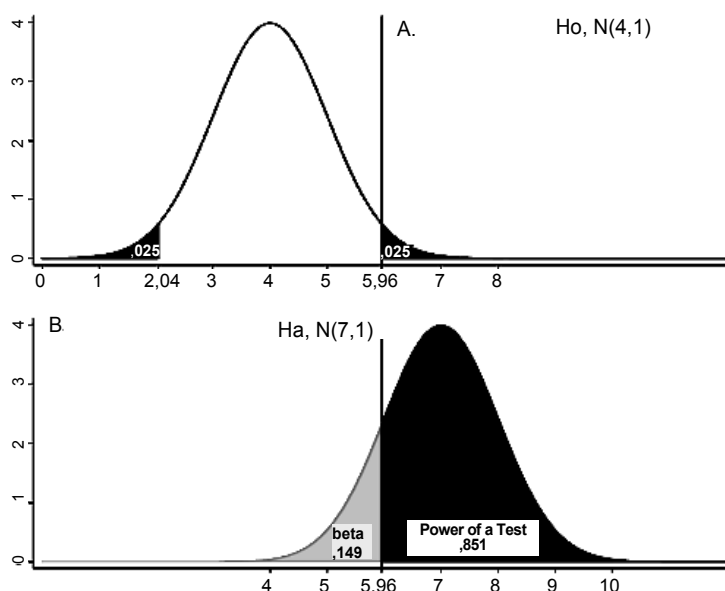
Khi tiến hành kiểm tra một đuôi với mức ý nghĩa 0,05, giá trị tới hạn là 1,725 có nghĩa là vùng bác bỏ giả thuyết từ 1,725 đến dương vô cùng và giá trị TS = 1,9 rơi vào vùng bác bỏ (Biểu đồ 3.4), có nghĩa là chúng ta bác bỏ giả thuyết H_0 . Giá trị P (diện tích dưới đường phân bố từ giá trị 1,9 đến dương vô cùng) là 0,036, nhỏ hơn so với mức ý nghĩa có nghĩa 0,05 có nghĩa là ít rủi ro hơn, do vậy chúng ta quyết định bác bỏ giả thuyết H_0 .



Biểu đồ 3.4. Giá trị tới hạn CV với kiểm tra một đuôi ở mức ý nghĩa 0,05 và độ tự do 20

3.1.4. Sai sót loại II và lực thống kê của phép kiểm tra

Lực thống kê là xác suất đưa ra một kết luận đúng về bác bỏ một giả thuyết H_0 sai (Greene 2000). Theo High (2000) lực thống kê là khả năng của một phép kiểm tra để phát hiện một ảnh hưởng, nếu ảnh hưởng đó thực sự tồn tại. Cohen (1988) cho rằng lực thống kê là xác suất dẫn đến một kết luận phản ánh sự tồn tại của một hiện tượng. Chúng ta có thể sử dụng lực thống kê để xác định dung lượng mẫu cần thiết hoặc chúng ta tính lực thống kê của một kiểm tra từ một dung lượng mẫu và mức ý nghĩa cho trước.



Biểu đồ 3.5. Sai sót loại II và lực thống kê của một phép kiểm tra

Giả thuyết H_0 trong thực tế là sai mà ta lại chấp nhận nó, tức là chúng ta đã làm một quyết định sai. Đó chính là sai sót loại II.

Xem xét hai phân bố trên biểu đồ 3.5. Biểu đồ A là một phân bố xác suất với giả thuyết H_0 : $N(4,1)$ và biểu đồ B là một phân bố xác suất với giả thuyết H_0 : $N(7,1)$. Ở biểu đồ A, xác suất của sai sót loại I là hai phần bôi màu đen ở hai đuôi của sự phân bố, 0,025 và 0,025. Giá trị 5,96 là giá trị tới hạn CV với mức ý nghĩa 0,05. Giả sử thực tế phân bố thực là phân bố chuẩn với trung bình bằng 7 và phương sai bằng 1, nhưng chúng ta đã sai sót cho rằng phân bố $N(4,1)$ là phân bố đúng. Chúng ta tiếp tục xem xét phần bên trái của phân bố xác suất ở biểu đồ B (phần từ âm vô cùng đến 5,96), khi chúng ta có giá trị tính toán TS nhỏ hơn 5,96 chúng ta không bác bỏ giả thuyết H_0 (quyết định sai) bởi vì chúng ta đã cho rằng phân bố A mới là phân bố đúng. Nếu chúng ta có giá trị TS là 5, chúng ta không bác bỏ giả thuyết H_0 bởi vì 5 không quá xa từ giá trị trung bình (4). Tuy nhiên, đây là một quyết định sai, bởi vì phân bố thực là phân bố ở biểu đồ B $N(7,1)$ và giả thuyết H_0 nên bị bác bỏ do giá trị TS 5 là đủ cách xa 7 ở mức ý nghĩa 0,05.

Phần diện tích màu đen là phần mà không bác bỏ giả thuyết H_0 : $\mu = 4$, khi giả thuyết H_0 là một giả thuyết sai, hay đây chính là sai sót loại II, được ký hiệu bởi β (beta). Phần bên phải của phân bố B, từ giá trị 5,96 (là một giá trị tới hạn CV trong phân bố A nhưng là một giá trị cụ thể trong phân bố B) đến dương vô cùng là $1 - \beta$. Phần này ($1 - \beta$) chính là lực thống kê. Như đã đề cập ở trên, phần này chính là xác suất bác bỏ giả thuyết H_0 sai: $\mu = 4$ hay chính là phát hiện một ảnh hưởng mà nó thực sự tồn tại. Thông lệ một kiểm tra có lực thống kê lớn hơn 0,8 hay $\beta \leq 0,2$ được xem là chấp nhận được (Mazen, Hemmasi, Lewis 1985).

3.2. Phân tích lực thống kê và ví dụ về phân tích lực thống kê

3.2.1. *Mối quan hệ của các yếu tố với lực thống kê*

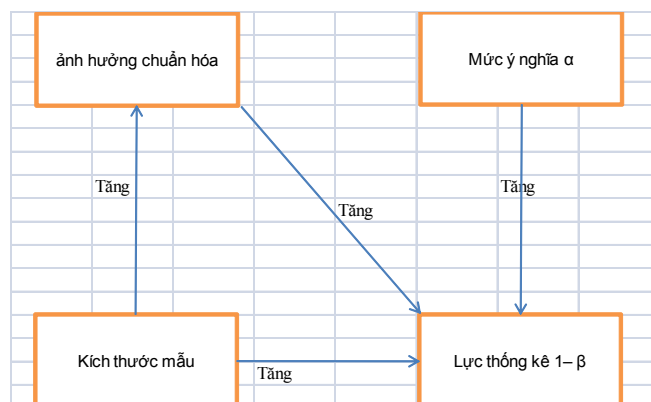
Khi nói đến phân tích lực thống kê chúng ta nói đến mối quan hệ giữa 4 hợp phần:

1. Kích thước ảnh hưởng chuẩn hóa (Standardized effect size): (i) kích thước ảnh hưởng = sự sai khác giữa giá trị của giả thuyết H_A và H_0 , (ii) mức biến động.
2. Kích thước mẫu (N) hay còn gọi là dung lượng mẫu.
3. Mức ý nghĩa α .
4. Lực thống kê $1 - \beta$.

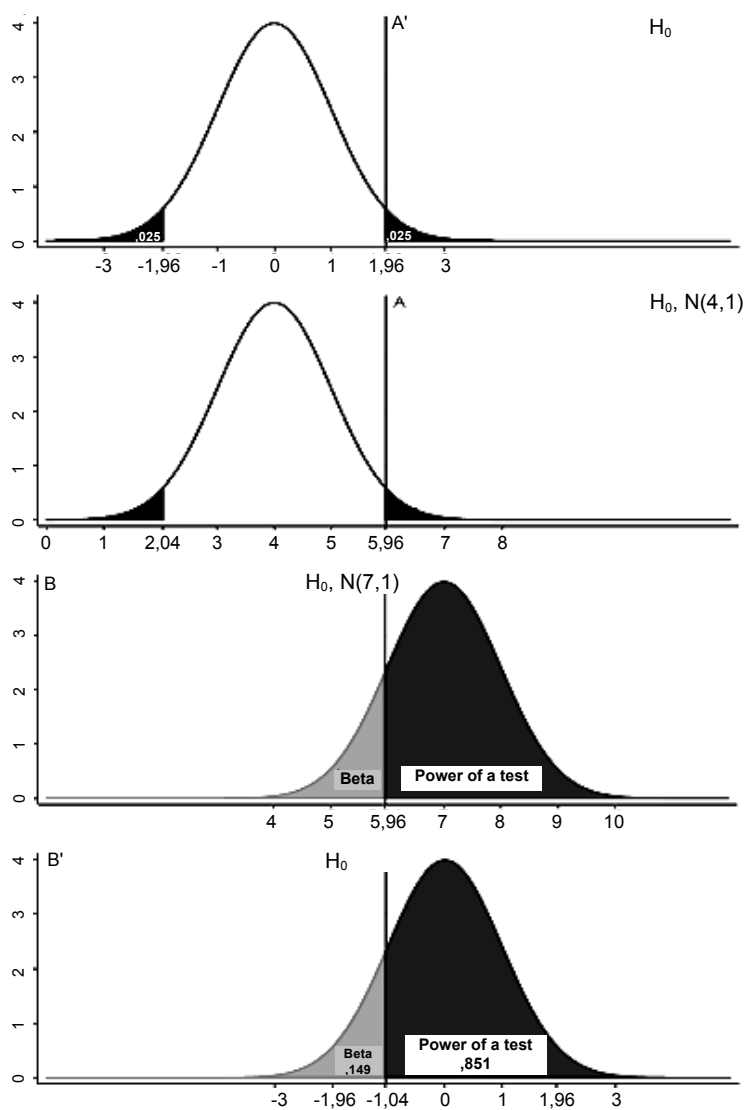
Kích thước ảnh hưởng chuẩn hóa, hay giá trị tính toán của các phép kiểm tra (ví dụ giá trị t hoặc giá trị F) được tính bằng cách kết hợp ảnh hưởng chưa chuẩn hóa và biến động/biến sai. Ví dụ trong kiểm tra t, ảnh hưởng chuẩn hóa là tỷ số giữa ảnh hưởng và chuẩn.

$$t = \frac{x - \mu}{se} \quad [3.1]$$

Biến động chính là độ lệch chuẩn hay sai số chuẩn của quần thể. Thông thường mức độ biến động này được lấy từ các nghiên cứu trước hay từ các nghiên cứu thí điểm hoặc chúng ta cần phải ước tính giá trị này.



Hình 3.1. Quan hệ giữa các thành phần trong phân tích lực thống kê



Biểu đồ 3.6. Sai sót loại I, II và lực thống kê

Khi tăng kích thước ảnh hưởng chuẩn hóa, lực thống kê tăng với điều kiện các thành phần khác là hằng số. Ảnh hưởng chuẩn hóa càng lớn có nghĩa rằng giá trị tham số mẫu khác nhiều so với giá trị giả thuyết, điều đó có nghĩa dễ dàng phát hiện ảnh hưởng. Ngược lại, nếu ảnh hưởng chuẩn hóa nhỏ thì khó phát hiện ảnh hưởng, mặc dù ảnh hưởng có tồn tại.

Có sự kết hợp/thỏa hiệp giữa sai sót loại I (α) và loại II (β). Khi chúng ta dịch chuyển đường thẳng đứng ở biểu đồ 3.6 qua bên trái có nghĩa chúng ta tăng sai sót loại I và đồng thời giảm sai sót loại II. Nếu chúng ta lựa chọn mức ý nghĩa 0,10, lực thống kê sẽ lớn (tương quan dương).

Khi dung lượng mẫu lớn, biến động (sai số chuẩn) sẽ nhỏ lại, do vậy kích thước ảnh hưởng chuẩn hóa lớn hơn. Ví dụ trong kiểm tra t, sai số chuẩn được tính là tỷ số giữa độ lệch chuẩn và căn bậc hai của mẫu.

Câu hỏi thường xuyên nhất khi tiến hành thí nghiệm là dung lượng mẫu nghiên cứu. Thành phần quan trọng nhất ảnh hưởng đến lực thống kê là kích thước mẫu. Trong thực tế chúng ta không thể thay đổi nhiều giá trị mức ý nghĩa, bởi vì theo thông lệ mức ý nghĩa lựa chọn là 0,05 hay 0,01. Tuy nhiên, khi số quan trắc quá lớn những ảnh hưởng bình thường sẽ có thể có ý nghĩa sai khác. Ngược lại nếu có quá ít quan sát thì kiểm tra giả thuyết sẽ yếu và tính thuyết phục sẽ thấp, do vậy cơ hội phát hiện ảnh hưởng là thấp mặc dù ảnh hưởng đó thực tế tồn tại (lực thống kê thấp).

3.2.2. Một số ví dụ về lực thống kê

Giả sử một biến ngẫu nhiên phân bố chuẩn với trung bình bằng 4 và phương sai bằng 1 (hình A biểu đồ 3.7). Giá trị tới hạn dương trong kiểm tra hai đuôi với mức ý nghĩa 0,05 là 5,96 bởi vì $1,96 = (5,96 - 4)/1$ là giá trị tới hạn tương ứng trong phân bố chuẩn hóa (hình A' trong biểu đồ 3.7). Giả thuyết H_0 của kiểm tra z $H_0: \mu = 4$. Hai vùng phản bác của giả thuyết này là từ âm vô cùng đến 2,04 và từ 5,96 đến dương vô cùng (hình A biểu đồ 3.7). Giá trị chuẩn hóa z được tính theo công thức [3.2]:

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{5,96 - 4}{1} = 1,96 \quad [3.2]$$

Giả sử phân bố thực là $N(7,1)$ - hình B - chứ không phải là $N(4,1)$. Nói một cách khác chúng ta đã sai khi cho rằng biến ngẫu nhiên tuân theo phân bố chuẩn với trung bình bằng 4. Xem xét hình B, phần diện tích dưới đường phân bố từ giá trị âm vô cùng đến 5,96; giá trị này chính là giá trị tới hạn sai trong hình A. Diện tích này cũng chính là sai sót loại II, cụ thể hơn là xác suất mà chúng ta không bác bỏ $H_0: \mu = 4$ khi mà giả thuyết $H_A: \mu = 7$ là giả thuyết đúng.

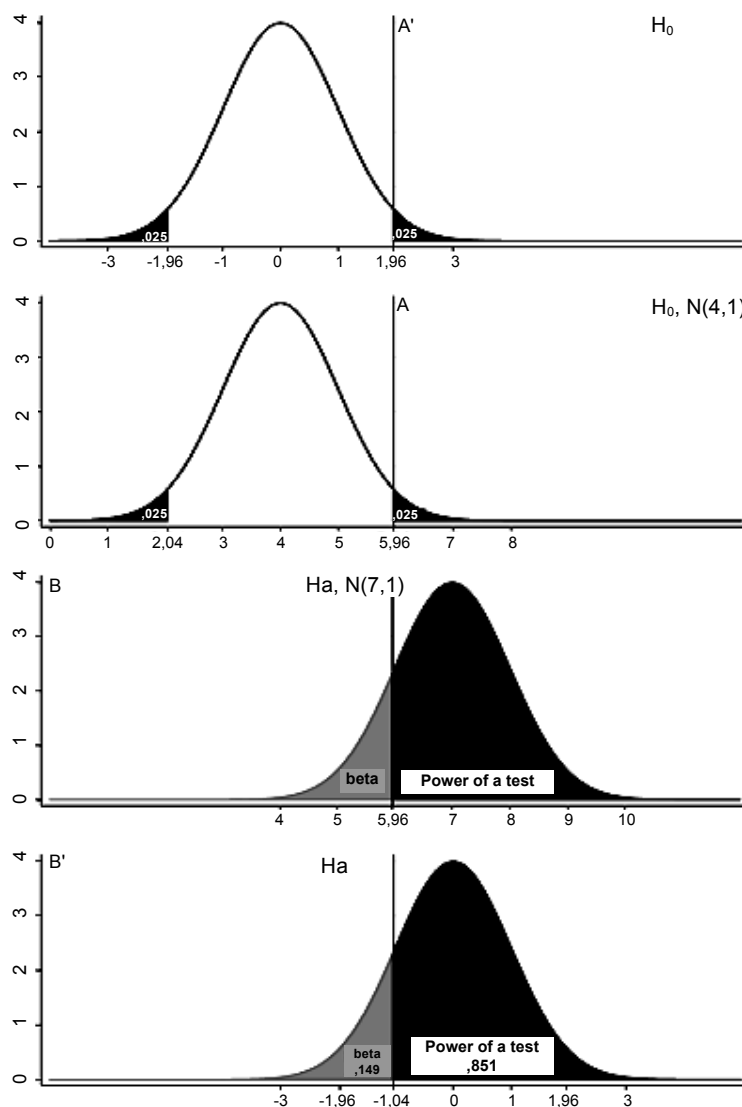
Giả sử chúng ta có trung bình mẫu bằng 4,5, giá trị z chuẩn hóa tương ứng là $0,5 = (4,5 - 4)/1 < \text{giá trị CV } 1,96$. Do vậy chúng ta không bác bỏ giả thuyết H_0 ở mức ý nghĩa 0,05. Nếu phân bố khác là $N(7,1)$, giá trị z chuẩn hóa là $2,5 = |(4,5 - 7)/1|$, giá trị

này lớn hơn giá trị CV 1,96. Do vậy, chúng ta bác bỏ giả thuyết $H_0: \mu = 7$ tại mức ý nghĩa 0,05.

Để tính sai sót loại II chúng ta chuyển giá trị 5,96 sang giá trị z

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{5,96 - 7}{1} = -1,04 \quad [3.3]$$

Sau đó tính diện tích từ âm vô cùng đến giá trị - 1,04 trong phân bố chuẩn hóa. Tra bảng phân bố xác suất ta có phần diện tích đó là 0,1492 (hình B' hay biểu đồ 3.8). Đây là phần xác suất mà chúng ta không bác bỏ giả thuyết H_0 khi H_0 sai. Như vậy nếu phân bố thực là $N(7,1)$, chúng ta có sai sót loại II khoảng 15%.



Biểu đồ 3.7. Lực thống kê của kiểm tra Z

Thông thường chúng ta muốn biết mức độ của một phép kiểm tra có thể phát hiện các ảnh hưởng nếu nó tồn tại, nói cách khác là lực thống kê của kiểm tra z như thế nào? Trong ví dụ này lực thống kê $= 1 - \beta = 0,851$. Nói một cách khác, trong 100 lần thử chúng ta có thể có 85 lần có thể phát hiện ảnh hưởng.

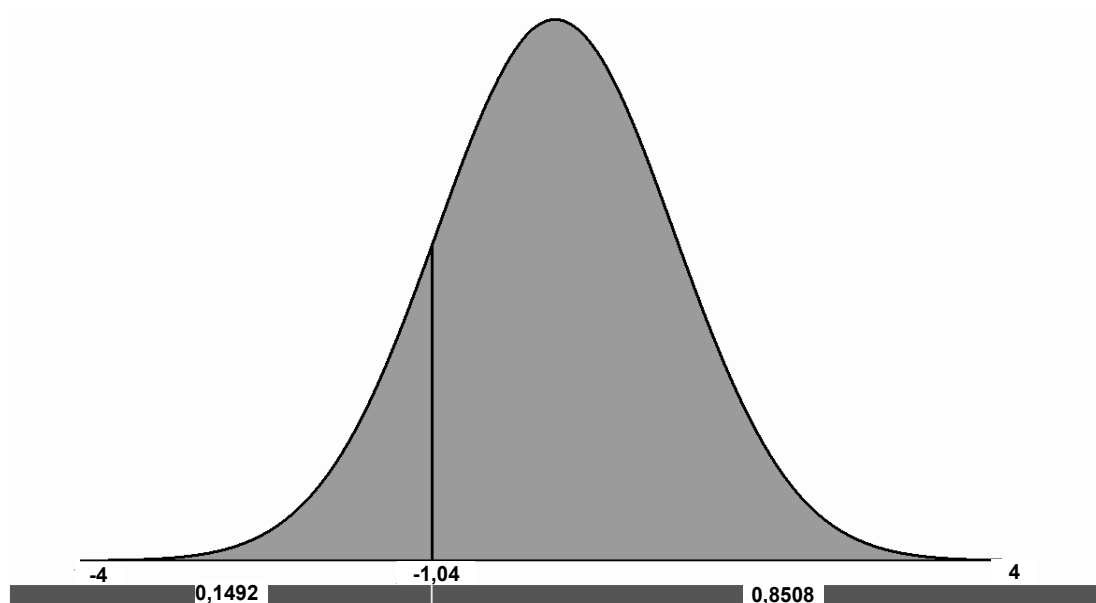
Tóm lại để tính lực thống kê, chúng ta có thể theo các bước sau:

Bước 1: Xác định giá trị tới hạn CV của một phân bố với mức ý nghĩa cho trước (căn cứ vào giả thuyết H_0). Trong ví dụ trên là, 2,04 và 5,96 (hình A và A' trong biểu đồ 3.7).

Bước 2: Sử dụng giá trị CV trong bước 1 để tính giá trị chuẩn hóa trong phân bố mới (hình B và B' ở biểu đồ 3.7). Chúng ta có giá trị $-1,04 = (5,96 - 7)/1$.

Bước 3: Sử dụng giá trị tính được ở bước 2 để tra giá trị β trong bảng phân bố xác suất, đó chính là sai sót loại II. Ở ví dụ trên $\beta = 0,149$.

Bước 4: Tính lực thống kê bằng cách áp dụng công thức $1 - \beta$. Trong ví dụ trên lực thống kê bằng 0,851.



Biểu đồ 3.8. Sai sót loại II và lực thống kê

3.3. Dung lượng mẫu cần thiết

Dung lượng mẫu nghiên cứu có ý nghĩa quan trọng liên quan đến độ chính xác của ước tính các giá trị trung bình nghiệm thức và lực thống kê của phép kiểm tra. Trong thực tế dung lượng mẫu nghiên cứu thường bị lãng quên. Sự hạn chế về kinh phí nghiên cứu thường là yếu tố hạn chế số lần lặp lại hay dung lượng mẫu nghiên cứu. Do vậy, thông thường dung lượng mẫu nghiên cứu được quyết định bởi điều

kiện kinh phí nghiên cứu. Xác định dung lượng mẫu là phép ước tính. Biến sai về tính trạng nghiên cứu có ý nghĩa lớn đến phép ước tính. Tuy nhiên, biến sai sử dụng cho phép ước tính là biến sai từ thí nghiệm khác chứ không phải là biến sai của thí nghiệm hiện tại.

3.3.1. Dung lượng mẫu trong kiểm tra sự khác nhau giữa 2 trung bình nghiệm thức

Phương pháp xác định số lần lặp lại thường dựa vào một phép kiểm tra giả thuyết về sự khác nhau giữa các trung bình nghiệm thức. Để hiểu hơn về phương pháp xác định số lần lặp lại chúng ta hãy xem xét một thí nghiệm với 2 nghiệm thức, thể hiện bởi 2 mẫu độc lập. Phương pháp ước tính số lần lặp lại dựa vào kiểm tra giả thuyết về sự khác nhau giữa 2 giá trị trung bình nghiệm thức $d = m_i - m_j$, với phương sai ngẫu nhiên là s^2 . Số lần lặp lại phụ thuộc vào:

- Phương sai (σ^2)
- Sự sai khác (mà được xem là có ý nghĩa trong thực tiễn) giữa hai trung bình (δ)
- Mức ý nghĩa (α) hay sai sót loại I
- Sai sót loại II (β)
- Lực thống kê $1 - \beta$ hay chính là xác suất để phát hiện δ

Trong trường hợp kiểm tra hai đuôi, số lần lặp lại cho mỗi nghiệm thức (r) được ước tính theo công thức sau:

$$r \geq 2 \left[z_{\alpha/2} + z_{\beta} \right]^2 \left(\frac{\sigma}{\delta} \right)^2 \quad [3.4]$$

Trong đó

- z là giá trị chuẩn hóa (tra bảng phân bố chuẩn hóa, $P(Z \geq z) = \alpha$)
- r là số lần lặp lại
- σ là độ lệch chuẩn
- δ là sự sai khác giữa hai giá trị trung bình

Chúng ta cũng có thể ước tính số lần lặp lại dựa vào hệ số biến dị (CV%). Thay thế σ trong công thức [3.4] bằng hệ số biến dị theo công thức $CV\% = 100(\sigma/\mu)$. Sự sai khác giữa hai trung bình nghiệm thức cũng phải được biểu diễn dưới dạng tỷ lệ % so với trung bình quần thể, $\delta\% = 100(\delta/\mu)$. Nhìn chung số lần lặp lại tăng nếu:

- Biến sai (%CV, σ , hay σ^2) tăng
- Sự sai khác nhau giữa 2 trung bình của 2 nghiệm thức (% δ hay δ) giảm
- Sai sót loại I giảm
- Lực thống kê tăng

Bảng 3.3. Số lần lặp lại cho các giá trị CV%, α , β và δ % khác nhau
(trong trường hợp kiểm tra hai đuôi sự khác nhau giữa 2 trung bình nghiệm thức)

		$\alpha = 0,05$		$\alpha = 0,01$	
		%δ			
CV%	1 - β	10	20	10	20
5	0,8	4	1	6	3
	0,95	7	2	9	3
10	0,8	16	4	24	6
	0,95	26	7	36	9

Một phương pháp tiếp cận khác để xác định số lần lặp lại cần thiết trong trường hợp thí nghiệm kiểm tra sự sai khác giữa hai giá trị trung bình là căn cứ vào các kiểm tra *Post hoc*, ví dụ kiểm tra Fisher LSD.

$$LSD = t_{\alpha/2, df_E} \sqrt{\frac{2\sigma^2}{n}} \quad [3.6]$$

Trong đó:

- LSD là giá trị sai khác ít nhất có ý nghĩa giữa hai giá trị trung bình của hai nghiệm thức
- t là giá trị quyết định tra bảng phân bố t - student
- α là sai sót loại I
- df_E là độ tự do của sai số ngẫu nhiên
- σ^2 là ước tính của phương sai hay chính là MS_E
- n là số lần lặp lại

Từ công thức [3.6] ta có

$$n = \frac{2t^2\sigma^2}{LSD^2} \quad [3.7]$$

Để đơn giản, chúng ta chuyển giá trị LSD và σ về cùng đơn vị thông qua tham số %CV như đã trình bày ở trên.

Xác định biến sai thông qua các tham số (%CV, σ) thông thường không đơn giản, chúng ta có thể dựa vào các kết quả nghiên cứu trước đây (dựa vào các thí nghiệm không thành công). Nhiều khi chúng ta chấp nhận kết quả gần đúng như sau:

- Các tính trạng sinh trưởng của động vật có vú có %CV khoảng 12%
- Các tính trạng sinh sản, %CV khoảng 20 - 40%
- Các tính trạng sản xuất, CV khoảng 20 - 25%

Điều này cũng có nghĩa bản chất của tính trạng nghiên cứu có ảnh hưởng lớn đến số lần lặp lại.

3.3.2. Dung lượng mẫu trong kiểm tra sự khác nhau giữa trung bình các nghiệm thức

Để kiểm tra sự khác nhau giữa các trung bình nghiệm thức chúng ta có thể sử dụng kiểm tra F (Phân tích ANOVA). Phần này đề cập đến phương pháp xác định số lần lặp lại trong phép kiểm tra F.

Ta có:

$$F_{\text{tính toán}} = MS_B / MS_E \quad [3.8]$$

Trong đó:

- $F_{\text{tính toán}}$: Giá trị F tính toán từ số liệu thí nghiệm
- MS_B : Trung bình bình phương giữa các nhóm còn gọi là trung bình bình phương của nhân tố nghiên cứu
- MS_E : Trung bình bình phương trong các nhóm còn gọi là trung bình bình phương ngẫu nhiên

Giả thuyết H_0 : $T_i = 0$ hay $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_i$

Lực thống kê trong kiểm tra F là $1 - \beta = P(F > F_{\alpha, v_1, v_2})$, trong đó v_1 là độ tự do tử số hay độ tự do giữa các nhóm; v_2 là độ tự do mẫu số hay độ tự do trong các nhóm. Lực thống kê cho kiểm tra F đã được tính sẵn với các giá trị α, v_1, v_2 và hàm Φ (công thức 3.9).

$$\Phi = \sqrt{\frac{r \sum_{i=1}^t T_i^2}{t\sigma^2}} \quad [3.9]$$

Trong đó:

- r là số lần lặp lại
- t là số nghiệm thức
- σ^2 là ước tính phương sai ngẫu nhiên $= MS_E$
- T_i là chênh lệch giữa trung bình nghiệm thức so với trung bình của các trung bình nghiệm thức

Ví dụ 3.1. Một nhóm nghiên cứu muốn tiến hành một kiểm tra với lực thống kê bằng 0,9, sai sót loại I $= 0,05$; $t = 3$; $MS_E = 0,22$

$$T_1 = \mu_1 - \mu_i = 0,8 - 0,3 = 0,5$$

$$T_2 = \mu_2 - \mu_i = 0,1 - 0,3 = -0,2$$

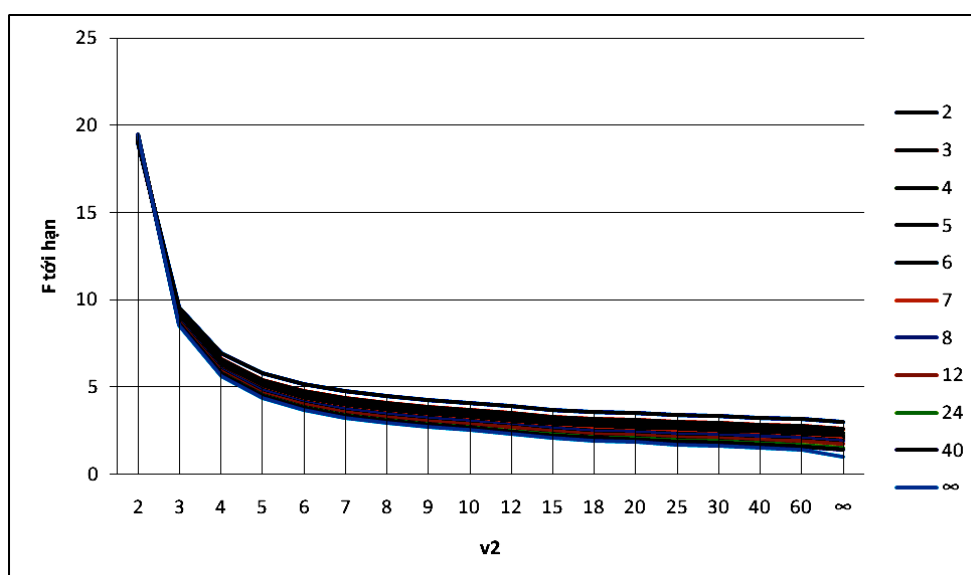
$$T_3 = \mu_3 - \mu_i = 0,0 - 0,3 = -0,3$$

$$\text{Nhu vậy } \sum_{i=1}^t T_i^2 = 0,38$$

Từ đó ta có: $\Phi^2 = \frac{r(0,38)}{3 \times 0,22} = r(0,58)$

Như vậy ta có $v_1 = 3 - 1 = 2$; và $v_2 = 3(r - 1)$. Giả sử ban đầu ta lựa chọn $r = 5$ khi đó ta có $v_2 = 12$ và giá trị $\Phi = 1,7$. Từ phụ lục 8 ta tra được lực thống kê khoảng 0,65, thấp hơn 0,9. Ta thử tăng lên $r = 9$ khi đó ta có $v_2 = 24$. Khi đó $\Phi = 2,3$, tra từ phụ lục 8 ta có lực thống kê lớn hơn 0,9. Như vậy để đạt được lực thống kê = 0,9, trong mỗi nghiệm thức cần ít nhất 9 lần lặp lại.

Một phương pháp tiếp cận khác liên quan đến xác định số lần lặp lại cần thiết trong trường hợp so sánh sự khác nhau giữa các giá trị trung bình là căn cứ vào độ tự do của ngẫu nhiên. Như đã đề cập ở trên, để kiểm tra sự khác nhau giữa các giá trị trung bình chúng ta có thể dựa vào kiểm tra F.



Biểu đồ 3.9. Giá trị F tới hạn của phân bố F với bậc tự do v_1, v_2 khi $\alpha = 0,05$

Qua biểu đồ trên ta thấy rằng giá trị F ít thay đổi khi độ tự do của ngẫu nhiên lớn hơn 20. Chúng ta có thể xem 20 như là giá trị tối thiểu của độ tự do của sai số ngẫu nhiên.

Ví dụ 3.2. Một thiết kế thí nghiệm được thiết kế theo kiểu hoàn toàn ngẫu nhiên (CRD^2) để nghiên cứu ảnh hưởng của 4 mức NSP trong khẩu phần đến phát xạ ammonia từ phân lợn, chúng ta cần ít nhất bao nhiêu lần lặp lại cho mỗi nghiệm thức. Chúng ta có df ngẫu nhiên trong kiểu thiết kế $CRD = t(r - 1)$; trong đó t là số nghiệm thức, r là số lần lặp lại. $t(r - 1) \geq 20$ hay $4r \geq 24$, như vậy $r \geq 6$.

² Xem chương 6 (Thiết kế thí nghiệm) để có thông tin về kiểu thiết kế này

Nguyên lý này cũng có thể được dùng để xác định số khối ít nhất trong kiểu thiết kế thí nghiệm khối ngẫu nhiên đầy đủ (RCB). **Ví dụ 3.3.** Cũng như ví dụ 3.2 nhưng do khối lượng lợn khi bắt đầu thí nghiệm khác nhau cho nên chúng ta phải thiết kế theo kiểu RCB để khống chế sự sai khác ban đầu. Cần ít nhất bao nhiêu khối? ta có đf ngẫu nhiên trong kiểu thiết kế $RCB = (b - 1)(t - 1) \geq 20$ hay $(b - 1) \times 3 \geq 20$; $b - 1 \geq 7$ hay $b \geq 6$.

Bảng 3.4. Số lần lặp lại đối với số liệu phân nhóm (2 nhóm); kiểm tra một đuôi

Tỷ lệ thành công thấp, %	$\bar{d}$ = Tỷ lệ thành công cao - Tỷ lệ thành công thấp													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
5	330	105	55	35	25	20	16	13	11	9	8	7	6	6
	460	145	76	48	34	26	21	17	15	13	11	9	8	7
	850	270	140	89	63	47	37	30	25	21	19	17	14	13
10	540	155	76	47	32	23	19	15	13	11	9	8	7	6
	740	210	105	64	44	33	25	21	17	14	12	11	9	8
	1370	390	195	120	81	60	46	37	30	25	21	19	16	14
15	710	200	94	56	38	27	21	17	14	12	10	8	7	6
	990	270	130	77	52	38	29	22	19	16	13	10	10	8
	1820	500	240	145	96	69	52	41	33	27	22	20	17	14
20	860	230	110	63	42	30	22	18	15	12	10	8	7	6
	1190	320	150	88	58	41	31	24	20	16	14	11	10	8
	2190	590	280	160	105	76	57	44	35	28	23	20	17	14
25	980	260	120	69	45	32	24	19	15	12	10	8	7	
	1360	360	165	96	63	44	33	25	21	16	14	11	9	
	2510	660	300	175	115	81	60	46	36	29	23	20	16	
30	1080	280	130	73	47	33	24	19	15	12	10	8		
	1500	390	175	100	65	46	33	25	21	16	13	11		
	2760	720	330	185	120	84	61	47	36	28	22	19		
35	1160	300	135	75	48	33	24	19	15	12	10	8		
	1600	410	185	105	67	46	33	25	20	16	12			
	2960	750	340	190	125	85	61	46	35	27	21			
40	1210	310	135	76	48	33	24	18	14	11				
	1670	420	190	105	67	46	33	24	19	14				
	3080	780	350	195	125	84	60	44	33	25				
45	1230	310	135	75	47	32	22	17	13					
	1710	430	190	105	65	44	31	22	17					
	3140	790	350	190	120	81	57	41	30					
50	1230	310	135	73	45	30	21	15						
	1710	420	185	100	63	41	29	21						
	3140	780	340	185	115	76	52	37						

Chú ý: Các con số ở hàng trên là trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,8$
 Các con số ở hàng giữa là trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,9$
 Các con số ở hàng dưới là trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,95$

3.3.3. Dung lượng mẫu cần thiết cho các số liệu phân thành hai nhóm

Trong một số nghiên cứu, đặc biệt trong lĩnh vực thú y, các số liệu thường được phân nhóm ví dụ, “sống” và “chết”; “mắc bệnh” và “không mắc bệnh”; “cải thiện” và “không cải thiện”.

Ví dụ 3.4. Để điều trị bệnh A trên lợn, bác sĩ thú y thường dùng thuốc B và tỷ lệ thành công thông thường khoảng 30%, gần đây xuất hiện một loại thuốc C. Nhóm bác sĩ thú y muốn so sánh tỷ lệ điều trị thành công của thuốc C và thuốc B. Nếu tỷ lệ điều trị thành công của thuốc C đạt khoảng 40% thì được cho là có ý nghĩa (về mặt khoa học hay kinh tế). Sai sót loại I của phép kiểm tra này là 0,05 và lực thống kê là 0,8. Cần bao nhiêu lợn bị bệnh cho nghiên cứu này?

Để xác định dung lượng mẫu cần thiết, trong trường hợp này chúng ta sử dụng bảng tra sẵn:

P_1 = Tỷ lệ thành công thấp = 30. δ = Sự sai khác giữa hai tỷ lệ = Tỷ lệ thành công cao - Tỷ lệ thành công thấp = 40 - 30 = 10. Trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,8$ chúng ta có $r = 360$. Như vậy tổng số mẫu nghiên cứu sẽ là 720.

Trong bảng 3.4 và 3.5 tỷ lệ P_1 cao nhất là 50%, nếu tỷ lệ thành công ở hai nhóm cần so sánh vượt quá thì khi tra bảng chúng ta dùng tỷ lệ không thành công.

Ví dụ 3.5. Để điều trị bệnh A trên lợn, bác sĩ thú y thường dùng thuốc B và tỷ lệ thành công thông thường khoảng 75%, gần đây xuất hiện một loại thuốc C. Nhóm bác sĩ thú y muốn so sánh tỷ lệ điều trị thành công của thuốc C và thuốc B. Nếu tỷ lệ điều trị thành công của thuốc C đạt khoảng 85% thì được cho là có ý nghĩa. Kiểm tra một đuôi với $\alpha = 0,05$ và lực thống kê là 0,9. Cần bao nhiêu lợn bị bệnh cho nghiên cứu này?

P_1 = Tỷ lệ không thành công là 15. δ = Sự sai khác giữa hai tỷ lệ = Tỷ lệ thành công cao - Tỷ lệ thành công thấp = 85 - 75 = 10. Trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,9$ chúng ta có $r = 270$ (bảng 3.4). Như vậy tổng số mẫu nghiên cứu sẽ là 540.

Bảng 3.5. Số lần lặp lại đối với số liệu phân nhóm (2 nhóm); kiểm tra hai đuôi

Tỷ lệ thành công thấp	δ = Tỷ lệ thành công cao - Tỷ lệ thành công thấp													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
5	420	130	69	44	31	24	20	16	14	12	10	9	9	7
	570	175	93	59	42	32	25	21	18	15	13	11	10	9
	960	300	155	100	71	54	42	34	28	24	21	19	16	14
10	680	195	96	59	41	39	23	19	16	13	11	10	9	7
	910	260	130	79	54	40	31	25	21	18	15	13	11	10
	1550	440	220	135	92	68	52	41	34	28	23	21	18	15

Tỷ lệ thành công thấp	$\bar{d} = \text{Tỷ lệ thành công cao} - \text{Tỷ lệ thành công thấp}$													
	5	10	15	20	25	30	35	40	45	50	55	60	65	70
15	910	250	120	71	48	34	26	21	17	14	12	10	9	8
	1220	330	160	95	64	46	35	27	22	19	16	13	11	10
	2060	560	270	160	110	78	59	47	37	31	25	21	19	16
20	1090	290	135	80	53	38	28	22	18	15	13	10	9	7
	1460	390	185	105	71	51	38	29	23	20	16	14	11	10
	2740	660	310	180	120	86	64	50	40	32	26	21	19	15
25	1250	330	150	88	57	40	30	23	19	15	13	10	9	
	1680	440	200	115	77	54	40	31	24	20	16	13	11	
	2840	740	340	200	130	92	68	52	41	32	26	21	18	
30	1380	360	160	93	60	42	31	23	19	15	12	10		
	1840	480	220	125	80	56	41	31	24	20	16	13		
	3120	810	370	210	135	95	69	53	41	32	25	21		
35	1470	380	170	96	61	42	31	23	18	14	11			
	1970	500	225	130	82	57	41	31	23	19	15			
	3340	850	380	215	140	96	69	52	40	31	23			
40	1530	390	175	97	61	42	30	22	17	13				
	2050	520	230	130	82	56	40	29	22	18				
	3340	850	380	215	140	95	68	52	50	28				
45	1560	390	175	96	60	40	28	21	16					
	2100	520	230	130	80	54	38	27	21					
	3550	890	390	215	135	92	64	47	34					
50	1560	390	170	93	57	38	26	19						
	2100	520	225	125	77	51	35	24						
	3550	880	380	210	130	86	59	41						

Chú ý: Các con số ở hàng trên là trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,8$
 Các con số ở hàng giữa là trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,9$
 Các con số ở hàng dưới là trong trường hợp $\alpha = 0,05$ và $1 - \beta = 0,95$

Hầu hết các phần mềm thống kê hiện tại đều có các trình ứng dụng để ước tính số lần lặp lại cần thiết có các thiết kế thí nghiệm khác nhau và cho các điều kiện cho trước về lực thống kê, sai sót loại I.... cho trước.

3.4. Bài tập

1. Giải thích ý nghĩa của sai sót loại I trong việc chấp nhận hay bác bỏ giả thuyết H_0 ?
2. Tại sao hạn chế sử dụng các cụm từ “có ý nghĩa”, “rất có ý nghĩa” khi suy diễn sự sai khác giữa hai hoặc nhiều hơn hai giá trị trung bình?
3. Một nhóm nghiên cứu ảnh hưởng của 3 mức protein thô 12, 15 và 17% trong khẩu phần đến tăng trọng của lợn. Giả thuyết H_0 của nhóm là $\bar{X}_{12\%} = \bar{X}_{15\%} = \bar{X}_{17\%}$. Hãy bình luận về cách đặt giả thuyết H_0 này?
4. Kết quả phân tích phương sai về sự khác nhau của khối lượng sơ sinh của dê Bách Thảo, Dê Boer và con lai F_1 (Boer $\times$ Bách Thảo), có giá trị $P = 0,06$. Hãy giải thích giá trị $P = 0,06$ có nghĩa là gì? Biện giải như thế nào sự khác nhau của khối lượng sơ sinh của 3 nhóm dê nêu trên khi $P = 0,06$?
5. Để điều trị bệnh A trên lợn, bác sĩ thú y thường dùng thuốc B và tỷ lệ thành công thông thường khoảng 20%, gần đây xuất hiện một loại thuốc C. Nhóm bác sĩ thú y muốn so sánh tỷ lệ điều trị thành công của thuốc C và thuốc B. Nếu tỷ lệ điều trị thành công của thuốc C đạt khoảng 35% thì được cho là có ý nghĩa (về mặt khoa học hay kinh tế). Sai sót loại I của phép kiểm tra này là 0,05 và lực thống kê là 0,9. Cần bao nhiêu lợn bị bệnh cho nghiên cứu này?
6. Một thiết kế thí nghiệm được thiết kế theo kiểu hoàn toàn ngẫu nhiên (CRD) để nghiên cứu ảnh hưởng của 3 mức thức ăn tinh trong khẩu phần đến tăng trọng của bò vàng Việt Nam trong giai đoạn vỗ béo, chúng ta cần ít nhất bao nhiêu lần lặp lại cho mỗi nghiệm thức?

Chương 4

MÔ HÌNH THỐNG KÊ VÀ PHÂN TÍCH PHƯƠNG SAI

Chương này trình bày hai nội dung chính: mô hình thống kê và phân tích phương sai. Đây là hai nội dung làm nền tảng cho thiết kế và phân tích số liệu thí nghiệm cũng như ước tính hiệu quả của các kiểu thiết kế thí nghiệm. Phần mô hình thống kê sẽ đề cập đến các chi tiết chính như thiết kế thí nghiệm và mô hình thống kê, ảnh hưởng cố định, ảnh hưởng ngẫu nhiên và các thành phần trong mô hình thống kê. Phần phân tích phương sai sẽ đề cập các chi tiết chính như lý do phân tích, các bước phân tích, trình bày và diễn giải kết quả phân tích phương sai, so sánh có kế hoạch, so sánh trực giao và phân tích hậu phương sai (Post hoc).

4.1. Mô hình thống kê

4.1.1. Mô hình thống kê cho thí nghiệm

Phân tích thống kê suy diễn được dựa vào các mô hình thống kê. Để diễn dịch kết quả phân tích đúng đòi hỏi phải hiểu mô hình thống kê. Mô hình thống kê có thể được hiểu một cách đơn giản là một phương trình gồm có hai vế: Vế bên phải tập hợp tất cả các nhân tố giải thích sự biến động của biến phụ thuộc, tuy nhiên ảnh hưởng của nhân tố ngẫu nhiên cần được không chế. Biến phụ thuộc nằm về phía bên trái của phương trình.

Các nhân tố giải thích sự biến động của biến phụ thuộc có thể phân chia thành hai nhóm: nhóm các nhân tố cố định và nhóm các yếu tố ngẫu nhiên. Ảnh hưởng cố định hay nhóm các yếu tố cố định là những yếu tố và các mức của yếu tố do người làm nghiên cứu quyết định và người nghiên cứu có thể không chế được sự thay đổi của nó. Ảnh hưởng của yếu tố cố định đến biến phụ thuộc là điều nhà nghiên cứu cần biết. **Ví dụ 4.1.** Trong thí nghiệm ảnh hưởng của mức protein thô trong khẩu phần (17%, 15% và 13%) đến phẩm chất thịt lợn lai F₁ (Móng Cái × Đại Bạch) thì yếu tố protein thô trong khẩu phần là yếu tố cố định.

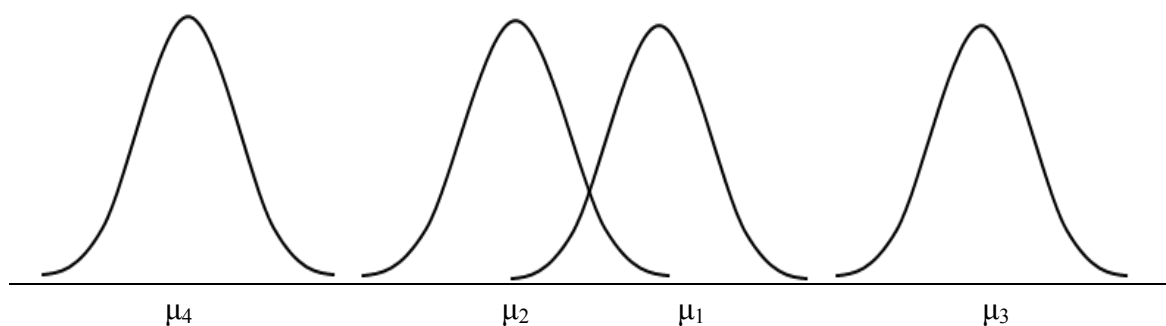
Ảnh hưởng ngẫu nhiên hay các nhân tố ngẫu nhiên là những yếu tố mà người nghiên cứu không quan tâm ảnh hưởng của nó đến biến phụ thuộc. Tuy nhiên ước tính phương sai của yếu tố ngẫu nhiên (MS_E hay MS_W) là tham số được nhà nghiên cứu quan tâm và tham số này chính là cơ sở xem xét hiệu quả của các kiểu thiết kế thí nghiệm. **Ví dụ 4.2** tiếp theo ví dụ 4.1: Sau khi kết thúc thí nghiệm 3 nhóm mẫu thịt được phân tích bởi 3 điều chế viên. Mỗi điều chế viên phân tích một nhóm mẫu. Mỗi nhóm mẫu thịt có đầy đủ các mẫu thịt từ lợn được ăn các khẩu phần với mức protein thô khác nhau. Cán bộ điều chế viên được xem là nhân tố ngẫu nhiên. Thông thường ảnh hưởng của kiểu thiết kế thí nghiệm được xem là ngẫu nhiên.

Trong các mô hình thống kê, biến độc lập có thể là biến định tính và cũng có thể là biến định lượng. Khi tất cả các biến độc lập là biến định tính thì mô hình phân tích ảnh hưởng của biến độc lập đến biến phụ thuộc gọi là mô hình phân tích phương sai (ANOVA). Ví dụ, mô hình phân tích ảnh hưởng của kiểu gen con bố và kiểu gen con mẹ đến khả năng sản xuất thịt của lợn đời con. Khi tất cả các biến độc lập là biến định lượng thì mô hình phân tích ảnh hưởng của biến độc lập đến biến phụ thuộc gọi là mô hình phân tích hồi quy (regression analysis model). Ví dụ mô hình phân tích ảnh hưởng của số đo vòng ngực và số đo dài thân chéo đến khối lượng của bò. Khi các biến độc lập bao gồm cả biến định tính và biến định lượng thì mô hình phân tích ảnh hưởng của biến độc lập đến biến phụ thuộc gọi là mô hình phân tích hỗn hợp (mixed model). Ví dụ, mô hình phân tích ảnh hưởng của kiểu gen lợn nái và chỉ số nhiệt ẩm đến khả năng sinh sản của lợn nái.

Các biến phụ thuộc có thể là các biến định lượng, ví dụ: tăng trọng hàng ngày (g/ngày) của lợn cũng có thể là các biến phân tử như sự chọn các giống bò khác nhau để nuôi sinh sản của người dân. Người dân có thể lựa chọn bò Vàng Việt Nam, bò Lai Red Sindhi, hay bò Brahman. Khi biến phụ thuộc là biến định lượng thì ta sử dụng các mô hình phân tích thường quy. Trong khi đó, khi biến phụ thuộc là biến phân tử thì chúng ta phải dùng các mô hình chuyên biệt như Tobit, Probit, Binomial hay Multinomial logit.

4.1.2. Mô hình đầy đủ và mô hình rút gọn và giả thuyết nghiên cứu

Mô hình thống kê đơn giản có thể được minh họa ở biểu đồ 4.1 với 4 quần thể của 4 nghiệm thức khác nhau. Giá trị của biến phụ thuộc của mỗi quần thể tuân theo quy luật phân bố chuẩn và giá trị trung bình là khác nhau giữa các quần thể. Các quần thể có các phương sai σ^2 được giả sử là giống nhau và không bị ảnh hưởng bởi các nghiệm thức. Hay nói cách khác phương sai quần thể của các nghiệm thức là đồng nhất.



Biểu đồ 4.1. Biểu đồ minh họa mô hình

Các quan trắc trong mỗi quần thể được biểu đạt dưới dạng tổng của các trung bình nghiệm thức và sai số thí nghiệm.

$$y_{ij} = \mu_i + e_{ij} \quad [4.1]$$

Trong đó:

- y_{ij} là giá trị quan trắc thứ j từ nghiệm thức i
- μ_i là trung bình các giá trị quan trắc của nghiệm thức i hay còn gọi trung bình nghiệm thức i
- e_{ij} là sai số thí nghiệm

Mô hình thống kê [4.1] là một ví dụ mô hình thống kê tuyến tính cho thí nghiệm một nhân tố/yếu tố thiết kế theo kiểu ngẫu nhiên hoàn toàn (CRD). Mỗi quan trắc chênh lệch với μ_i một lượng e_{ij} . Phương sai của sai số thí nghiệm σ^2 là phương sai của e_{ij} và được giả sử là giống nhau cho tất cả các nghiệm thức.

Mô hình thống kê cho thí nghiệm phản ánh mối quan hệ giữa các nghiệm thức và các quan trắc. Mô hình [4.1] còn được gọi là mô hình đầy đủ. Mô hình này bao gồm một trung bình riêng biệt cho mỗi quần thể nghiệm thức. Mô hình rút gọn có dạng:

$$y_{ij} = \mu + e_{ij} \quad [4.2]$$

Mô hình [4.2] có nghĩa là tất cả các quan trắc đều thuộc một quần thể giống nhau với giá trị trung bình μ . Mô hình [4.1] và [4.2] thể hiện hai loại giả thuyết thống kê. Mô hình rút gọn [4.2] thể hiện giả thuyết H_0 : Không có sự khác nhau giữa các trung bình của các nghiệm thức $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Mô hình đầy đủ [4.1] thể hiện giả thuyết H_A : Ít nhất trung bình của một nghiệm thức khác với một trung bình của nghiệm thức khác.

4.2. Phân tích phương sai

4.2.1. Lý do phân tích phương sai và điều kiện để phân tích phương sai

Để kiểm tra giả thuyết hay để quyết định chấp nhận mô hình đầy đủ [4.1] hay mô hình rút gọn [4.2], chúng ta có thể tiến hành phân tích phương sai. Hay cụ thể hơn để diễn dịch sự khác nhau giữa các trung bình của các nghiệm thức, chúng ta có thể sử dụng phép phân tích phương sai. Phân tích phương sai căn cứ vào sự biến động/biến sai do nhân tố thí nghiệm gây nên (biến sai giữa các nhóm hay biến sai giữa các mức của nhân tố thí nghiệm) và biến sai không phải do yếu tố thí nghiệm gây nên (biến sai trong các nhóm hay biến sai trong mỗi mức của nhân tố thí nghiệm) để diễn dịch sự khác nhau giữa các giá trị trung bình của các nghiệm thức. Nói cách khác là phân tích phương sai lấy tham số đặc trưng cho mức độ phân tán (phương sai) để diễn dịch sự khác nhau của tham số đặc trưng cho mức độ tập trung (trung bình).

Để diễn dịch sự khác nhau giữa các giá trị trung bình của các mẫu độc lập, chúng ta có thể dùng phép kiểm tra t . Phép kiểm tra t cho phép so sánh lần lượt sự khác nhau giữa hai giá trị trung bình. Nếu chúng ta muốn so sánh sự khác nhau giữa 3 giá trị trung bình chúng ta cần thực hiện 6 phép kiểm tra t , tương tự như vậy để so sánh sự khác nhau

giữa 4 giá trị trung bình chúng ta cần thực hiện 25 phép kiểm tra t. Số phép kiểm tra t sẽ tăng lên nhanh chóng khi số trung bình cần so sánh tăng lên. Như vậy, kiểm tra t có thể hao phí thời gian. Mặt khác thực hiện nhiều kiểm tra t làm tăng sai sót loại I. Sai sót loại I sẽ lớn hơn mức α của mỗi phép kiểm tra.

Phân tích ANOVA có thể khắc phục được nhược điểm của kiểm tra t bằng một phép kiểm tra đơn cho sự khác nhau tổng thể giữa toàn bộ các trung bình của các nghiệm thức, tại một mức α . Nếu H_0 bị bác bỏ thì có ít nhất một trung bình khác với một trung bình khác. Sau đó có thể thực hiện kiểm tra *Post hoc* hay so sánh có kế hoạch để quyết định các giá trị trung bình cụ thể nào là sai khác với các giá trị trung bình nào.

Để có thể tiến hành phân tích phương sai về sự khác nhau của các giá trị trung bình của các nghiệm thức, các quần thể cần phải đáp ứng hai điều kiện sau:

- Tất cả các quần thể có phân bố chuẩn.
- Phương sai của tất cả các quần thể là tương đương: $\sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_t$ (Giả thuyết về tính đồng nhất của phương sai). Trong đó t là số nghiệm thức hay chính là số mức của nhân tố thí nghiệm mà các giá trị trung bình của nó cần so sánh.

4.2.2. Các bước phân tích phương sai

4.2.2.1. Phân chia tổng biến động của một tập hợp số liệu

Tổng biến động của một tập hợp số liệu có thể được chia thành hai phần, một phần căn cứ vào sự biến động của các giá trị trong các mẫu và một phần căn cứ vào sự biến động giữa các trung bình mẫu. Để hiểu rõ tổng biến động và các biến động thành phần, chúng ta xem xét ví dụ 4.3.

Ví dụ 4.3. Ảnh hưởng của các mức protein thô (CP) trong khẩu phần (13%, 15% và 17%) đến tăng trọng (kg/tháng) của lợn lai F₁ (Móng Cái × Đại Bạch) được bộ môn Dinh dưỡng và Thức ăn, khoa Chăn nuôi Thú y, Đại học Nông lâm Huế nghiên cứu, kết quả được trình bày ở bảng 4.1. Thí nghiệm được tiến hành trên 24 con lợn (N), số lần lặp lại ở mỗi nghiệm thức CP là 8 con lợn (n). Giả sử tăng trọng của lợn ở mỗi nghiệm thức CP đáp ứng yêu cầu của phân tích phương sai.

Bảng 4.1. Tăng trọng (kg/tháng) của lợn lai F₁ (Móng Cái × Đại Bạch) được nuôi với 3 mức protein thô trong khẩu phần 13%, 15% và 17%

	Mức protein thô (CP)		
	13%	15%	17%
11	11	15	13
3	3	14	14
5	5	10	6
10	10	12	10
7	7	13	12
8	8	6	10,5
9	9	9	16
2	2	10	13

N thường được dùng để ký hiệu tổng số đơn vị thí nghiệm và n thường được ký hiệu số đơn vị thí nghiệm nhận cùng một nghiệm thức hay còn gọi là số lần lặp lại.

Từ bảng 4.1 ta tính trung bình tăng trọng của lợn ở 3 mức protein 13%, 15% và 17% lần lượt là 6,875; 11,125 và 11,813 kg/tháng. Trung bình của các nghiệm thức được ký hiệu là M. Như vậy ta có $M_1 = 6,875$; $M_2 = 11,125$ và $M_3 = 11,813$ tương ứng với 13%; 15% và 17% CP trong khẩu phần. Trung bình tăng trọng của toàn bộ các cá thể nghiên cứu, thường được ký hiệu là GM (còn gọi là trung bình tổng thể), là 9,938 kg/tháng.

Tổng các bình phương độ lệch giữa mỗi quan trắc so với trung bình tổng thể hay còn gọi là tổng bình phương tổng thể, ký hiệu SS_T bao gồm hai thành phần: (1) tổng các bình phương độ lệch giữa các giá trị quan trắc trong các nghiệm thức so với trung bình của nghiệm thức hay còn gọi là tổng các bình phương trong các nhóm, ký hiệu SS_W (SS_W còn được ký hiệu là SS_E) và (2) tổng các bình phương độ lệch giữa các trung bình nghiệm thức so với trung bình tổng thể hay còn gọi là tổng các bình phương giữa các nhóm, ký hiệu SS_B . Do vậy, chúng ta có thể viết:

$$SS_T = SS_W + SS_B \quad [4.3]$$

- *Tổng bình phương tổng thể*

Tổng bình phương tổng thể được tính theo công thức [4.4]

$$SS_T = \sum_{i=1}^N (X_i - GM)^2 \quad [4.4]$$

Trong đó: X_i là giá trị của các quan trắc; SS_T là tổng các bình phương tổng thể; GM là trung bình tổng thể và N là tổng số các quan trắc. Phương pháp tính toán SS_T và giá trị SS_T ở ví dụ 4.3 được trình bày ở bảng 4.2. Qua bảng 4.2 ta thấy rằng giá trị $SS_T = 314,16$.

Khi chúng ta tiến hành phân tích phương sai không có sự trợ giúp của máy vi tính, công thức sau nên được sử dụng.

$$SS_T = \sum X_i^2 - \frac{(\sum X_i)^2}{N} \quad [4.5]$$

Trong đó N là tổng số quan trắc.

- *Tổng các bình phương trong các nhóm*

Tổng các bình phương trong các nhóm được tính theo công thức [4.6].

$$SS_W = \Sigma(X_1 - M_1)^2 + \Sigma(X_2 - M_2)^2 + \dots + \Sigma(X_t - M_t)^2 \quad [4.6]$$

Bởi vì $\Sigma(X - M)^2$ trong bất cứ một nghiệm thức nào là tổng các bình phương độ lệch, chúng ta có thể viết: $SS_W = SS_1 + SS_2 + \dots + SS_t$. Trong đó t chính là số nghiệm thức hay chính là số mức của nhân tố nghiên cứu. Có nghĩa là để tính toán SS_W , chúng ta phải tính $\Sigma(X - M)^2$ trong mỗi nghiệm thức, sau đó cộng tất cả của t nghiệm thức.

Phương pháp tính toán SS_W và giá trị SS_W của ví dụ 4.3 được trình bày ở bảng 4.2. Qua bảng 4.2 ta thấy rằng giá trị $SS_W = 74,88 + 60,88 + 63,97 = 199,72$.

Khi phân tích phương sai không có sự trợ giúp của máy vi tính, chúng ta nên tính các giá trị $SS_1, SS_2, SS_3 \dots SS_t$ theo công thức [4.7]. Mỗi lần thực hiện trong mỗi nghiệm thức, sau đó tính $SS_W = SS_1 + SS_2 + \dots + SS_t$

$$SS = \sum X^2 - \frac{(\sum X)^2}{n} \quad [4.7]$$

- *Tổng các bình phương giữa các nhóm*

Tổng các bình phương giữa các nhóm được tính theo công thức [4.8]:

$$SS_B = n_1(M_1 - GM)^2 + n_2(M_2 - GM)^2 + \dots + n_t(M_t - GM)^2 \quad [4.8]$$

Để có được SS_B , chúng ta tính $\Sigma(M - GM)^2$ trong mỗi nghiệm thức, sau đó cộng tất cả lại với nhau. Kết quả tính toán được thể hiện ở cột (5) và (6) của bảng 4.2. Giá trị SS_B của ví dụ 4.3 = $75,03 + 11,28 + 28,13 = 114,44$.

Bảng 4.2. Cách tính tổng các bình phương của ví dụ 4.3

	X	(1) X - GM	(2) (X - GM) ²	(3) X - M	(4) (X - M) ²	(5) M - GM	(6) (M - GM) ²
Mức CP 13% n ₁ = 8 M ₁ = 6,875	11	1,06	1,13	4,13	17,02	- 3,06	9,38
	3	- 6,94	48,13	- 3,88	15,02	- 3,06	9,38
	5	- 4,94	24,38	- 1,88	3,52	- 3,06	9,38
	10	0,06	0,00	3,13	9,77	- 3,06	9,38
	7	- 2,94	8,63	0,13	0,02	- 3,06	9,38
	8	- 1,94	3,75	1,13	1,27	- 3,06	9,38
	9	- 0,94	0,88	2,13	4,52	- 3,06	9,38
	2	- 7,94	63,00	- 4,88	23,77	- 3,06	9,38
	Tổng theo nghiệm thức	149,91			74,88 = SS ₁		75,03 = 8(9,38) = n ₁ (M ₁ - GM) ²
Mức CP 15% n ₂ = 8 M ₂ = 11	15	5,06	25,63	3,88	15,02	1,19	1,41
	14	4,06	16,50	2,88	8,27	1,19	1,41
	10	0,06	0,00	- 1,13	1,27	1,19	1,41
	12	2,06	4,25	0,88	0,77	1,19	1,41
	13	3,06	9,38	1,88	3,52	1,19	1,41
	6	- 3,94	15,50	- 5,13	26,27	1,19	1,41
	9	- 0,94	0,88	- 2,13	4,52	1,19	1,41
	10	0,06	0,004	- 1,13	1,27	1,19	1,41
	Tổng theo nghiệm thức	72,16			60,88 = SS ₂		11,28 = 8(1,41) = n ₂ (M ₂ - GM) ²

	X	(1) X - GM	(2) (X - GM) ²	(3) X - M	(4) (X - M) ²	(5) M - GM	(6) (M - GM) ²
Mức CP 17% n ₃ = 8 M ₃ = 12	13	3,06	9,38	1,19	1,41	1,88	3,52
	14	4,06	16,50	2,19	4,79	1,88	3,52
	6	- 3,94	15,50	- 5,81	33,79	1,88	3,52
	10	0,06	0,00	- 1,81	3,29	1,88	3,52
	12	2,06	4,25	0,19	0,04	1,88	3,52
	10,5	0,56	0,32	- 1,31	1,72	1,88	3,52
	16	6,06	36,75	4,19	17,54	1,88	3,52
	13	3,06	9,38	1,19	1,41	1,88	3,52
Cộng tổng			92,09		63,97 = SS ₃		28,13 = 8(3,52) = n ₃ (M ₃ - GM) ²
Tổng chung			314,16 SS _T		199,72 SS _W		114,44 SS _B

Khi phân tích ANOVA không có sự trợ giúp của máy vi tính, chúng ta nên tính theo công thức [4.9].

$$SS_B = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \dots + \frac{(\sum X_t)^2}{n_t} - \frac{(\sum X_i)^2}{N} \quad [4.9]$$

Kết quả tổng các bình phương ở ví dụ 4.3 được tóm tắt ở bảng 4.3.

Bảng 4.3. Tổng các bình phương của ví dụ 4.3

SS _T	314,16
SS _W	199,72
SS _B	114,44

Như vậy $SS_T = SS_W + SS_B$. Điều này luôn luôn đúng cho mọi tập hợp số liệu. Do vậy, trong thực tế chúng ta không cần thiết phải tính cả 3 loại tổng các bình phương mà chỉ cần tính 2 trong 3 loại và thông thường SS_T và SS_W được tính. Khi đó $SS_B = SS_T - SS_W$.

4.2.2.2. Trung bình bình phương và độ tự do

Như đã nêu trong phần 4.2.1 phân tích phương sai dựa vào biến sai do nhân tố nghiên cứu tạo nên và biến sai do ngẫu nhiên gây nên, hay căn cứ vào phương sai trong các nghiệm thức và phương sai giữa các nghiệm thức. Phương sai trong các nghiệm thức còn gọi là trung bình bình phương trong các nghiệm thức (MS_W hay MS_E). Phương sai giữa các nghiệm thức còn gọi là trung bình bình phương giữa các nghiệm thức (MS_B).

- *Trung bình bình phương trong các nghiệm thức*

Khi tiến hành kiểm tra t cho 2 mẫu độc lập chúng ta phải ước tính một phương sai chung, ký hiệu là S_p^2 . Giá trị của phương sai chung chính là tỷ số giữa tổng chung của tổng các bình phương đối với hai mẫu nghiên cứu cho độ tự do chung. Nguyên tắc tương tự được áp dụng khi phân tích phương sai.

$$MS_w = \frac{SS_{\text{chung}}}{df_{\text{chung}}} = \frac{SS_1 + SS_2 + \dots + SS_t}{(n_1 - 1) + (n_2 - 1) + \dots + (n_t - 1)} = \frac{SS_w}{(n_1 + n_2 + \dots + n_t) - t} = \frac{SS_w}{N - t}$$

Độ tự do chung, $N - t$ gọi là độ tự do trong các nghiệm thức (df_w). Do vậy, bình phương trung bình trong các nhóm là tổng các bình phương trong các nhóm chia cho độ tự do trong các nhóm.

$$MS_w = \frac{SS_w}{df_w} \quad [4.10]$$

Trong ví dụ 4.3 nêu ở trên ta có $N = 24$ và $t = 3$, do vậy $df_w = 21$ và $SS_w = 199,72$ nên $MS_w = 199,72/21 = 9,51$.

Trong kiểm tra t, S_p^2 là phương sai ước tính của quần thể σ^2 . Tương tự như vậy, MS_w là một ước tính của phương sai quần thể, σ^2 . Do vậy, chúng ta có thể nói rằng các quan trắc trong ví dụ 4.3 xuất phát từ quần thể có giá trị phương sai ước tính 9,51.

- *Trung bình bình phương giữa các nghiệm thức*

Trung bình bình phương giữa các nghiệm thức được tính theo công thức [4.11].

$$MS_B = \frac{SS_B}{df_B} \quad [4.11]$$

Khi chúng ta chấp nhận giả thuyết H_0 , khi đó MS_w tương đương với MS_B , bởi vì khi chấp nhận H_0 , MS_B cũng là một ước tính của σ^2 , nhưng căn cứ vào sự biến động giữa các trung bình nghiệm thức hơn là sự biến động của các giá trị trong các nghiệm thức. Để có thể chứng minh điều này chúng ta phải vận dụng định lý giới hạn trung tâm như đã đề cập trong học phần xác suất thống kê.

Bởi vì SS_B của ví dụ 4.3 là 114,44 và $df_B = t - 1 = 2$ cho nên giá trị trung bình bình phương giữa các nghiệm thức là $114,44/2 = 57,22$.

Khi chúng ta bác bỏ giả thuyết H_0 có nghĩa là MS_B lớn một cách đáng kể so với phương sai của quần thể. Điều này cũng có nghĩa, trong trường hợp này các trung bình nghiệm thức biến động mạnh hơn so với biến động ngẫu nhiên, do vậy MS_B không phải là một ước tính của phương sai quần thể. Phương sai giữa các nghiệm thức lớn hơn so với phương sai quần thể.

- *Độ tự do*

Khi tính MS_B và MS_W , chúng ta đã sử dụng hai giá trị df_B và df_W . df_B còn được gọi là độ tự do giữa các nghiệm thức và df_W là độ tự do trong các nghiệm thức. df_B và df_W được xác định như sau:

$$df_W = N - t \quad [4.12]$$

$$df_B = t - 1 \quad [4.13]$$

Ngoài ra chúng ta còn có thể xác định độ tự do tổng số theo công thức 4.10.

$$df_T = N - 1 \quad [4.14]$$

Từ ba công thức 4.12; 4.13 và 4.14 thấy rằng:

$$N - t + t - 1 = N - 1 \text{ hay } df_W + df_B = df_T.$$

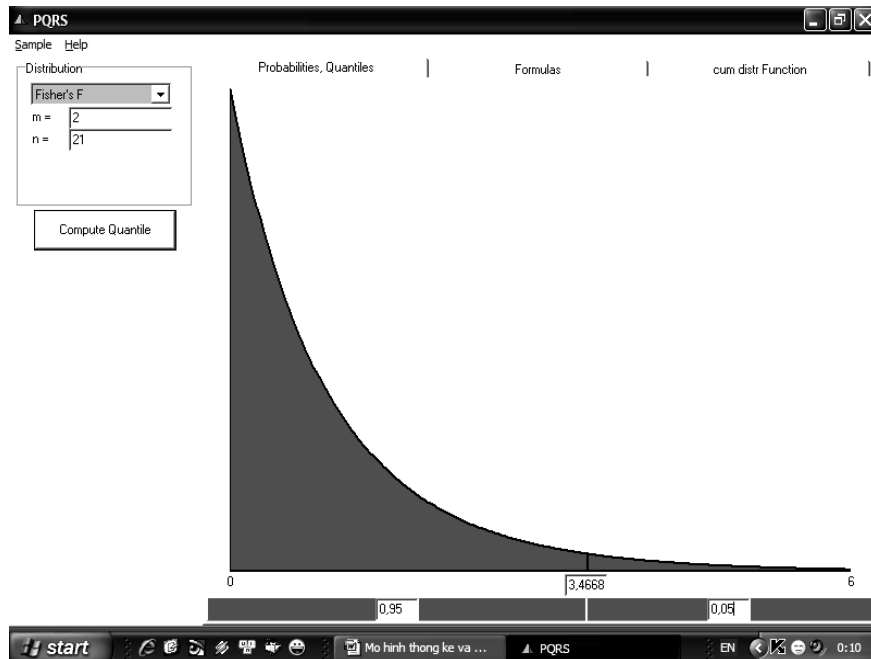
Điều này có nghĩa độ tự do tổng số được phân chia thành hai thành phần: độ tự do giữa các nghiệm thức và độ tự do trong các nghiệm thức.

4.2.2.3. Kiểm tra F và phân tích phương sai

Như đã đề cập ở trên, bản chất của phân tích phương sai là dựa trên giá trị của phương sai giữa các nghiệm thức và phương sai trong các nghiệm thức. Khi H_0 đúng thì hai loại phương sai đó tương đương nhau và khi H_A đúng thì hai loại phương sai đó khác nhau đủ lớn để kết luận nhân tố thí nghiệm có ảnh hưởng đến chỉ tiêu nghiên cứu. Để kiểm tra sự khác nhau giữa hai loại phương sai này chúng ta dựa vào hàm phân bố Fisher (F) và phép kiểm tra này gọi là kiểm tra F. Bản chất của phân bố F chính là sự phân bố của tỷ lệ của hai phương sai, ước tính phương sai 1 và ước tính phương sai 2. Giá trị ước tính của hai phương sai thay đổi tùy theo các mẫu khác nhau của quần thể, do vậy tỷ lệ của hai phương sai cũng thay đổi.

$$F = \frac{\text{ước tính phương sai 1}}{\text{ước tính phương sai 2}} \quad \text{nếu } \sigma_1^2 = \sigma_2^2 \quad [4.15]$$

Phân bố F được đặc trưng bởi hai độ tự do, độ tự do của ước tính phương sai 1 và độ tự do của ước tính phương sai 2. Một số tài liệu còn gọi là độ tự do tử số và độ tự do mẫu số. Cụ thể hơn cho trường hợp ví dụ 4.3 đó là độ tự do giữa các nghiệm thức và độ tự do trong các nghiệm thức. Hình dạng phân bố F với độ tự do 2 và 21 trong ví dụ 4.3 được thể hiện ở biểu đồ 4.2. Qua biểu đồ 4.2 ta thấy phân bố F là một phân bố lệch dương, điều này là hiển nhiên do tỷ lệ của hai phương sai luôn luôn dương.



Biểu đồ 4.2. Hình dạng của phân bố F với độ tự do 2 và 21
(Hình dạng phân bố được xác định bởi phần mềm PQRS)

Khi H_0 là đúng thì tỷ lệ của MS_B và MS_W là phân bố F với độ tự do tử số và độ tự do mẫu số tương ứng là df_B và df_W . Chúng ta tính kiểm tra thống kê.

$$F = \frac{MS_B}{MS_W} \quad [4.16]$$

Giá trị F được tính từ công thức 4.16 được gọi là F tính toán, hàm ý do người làm nghiên cứu tính toán mà có được. Cách gọi F tính toán nhằm phân biệt với giá trị F không do tính toán tạo nên mà do tra bảng phân bố Fisher căn cứ vào độ tự do tử số và độ tự do mẫu số mà có được. Nếu F tính toán nhỏ hơn 1 hoặc không lớn lắm thì giả thuyết H_0 được chấp nhận hay giả thuyết thay thế H_A bị bác bỏ. Điều này có nghĩa, biến động giữa các nghiệm thức có thể chỉ do yếu tố ngẫu nhiên chứ không phải do yếu tố thí nghiệm tạo nên.

Chúng ta chấp nhận giả thuyết H_A khi giá trị F tính toán đủ lớn. Hàm ý đủ lớn ở đây có nghĩa là khi so sánh với giá trị F tra bảng từ mức ý nghĩa α mà chúng ta đã chọn (như đã đề cập trong phần kiểm tra giả thuyết và lực thống kê ở chương trước) cùng với df_W và df_B . Giá trị F tính toán có thể được tra trong bảng F để xác định giá trị xác suất P. Bảng F thường nằm ở phần phụ lục của tất cả các sách liên quan đến xác suất thống kê và thiết kế thí nghiệm. Cột trong bảng F chính là độ tự do tử số hay độ tự do giữa các nghiệm thức (df_B). Hàng trong bảng F chính là độ tự do mẫu số hay độ tự do trong các nghiệm thức (df_W). Ngoài ra, chúng ta có thể tra giá trị F bằng cách sử dụng các phần

mềm thông dụng như PQRS, thậm chí phần mềm không chuyên dụng như Excel. Giả sử khi kiểm tra giả thuyết ở ví dụ 4.3 chúng ta chọn mức độ tin cậy $\alpha = 0,05$, khi đó giá trị F tra bảng ở độ tự do (2, 21) ở mức $\alpha = 0,05$ là 3,47.

Giá trị F tính toán theo công thức 4.16 sẽ là:

$$F_{\text{tính toán}} = \frac{MS_B}{MS_W} = \frac{57,22}{9,51} = 6,02$$

Như vậy, giá trị $F_{\text{tính toán}} > \text{giá trị } F_{\text{tra bảng}}$, có nghĩa là giả thuyết H_0 bị bác bỏ và giả thuyết H_A được chấp nhận. Chúng ta kết luận rằng các mức protein khác nhau trong khẩu phần có ảnh hưởng đến tăng trọng của lợn lai F_1 (Móng Cái $\times$ Đại Bạch). Để nâng cao tính thuyết phục của kết luận khi đã bác bỏ giả thuyết H_0 ở mức $\alpha = 0,05$ thì chúng ta nên tiếp tục kiểm tra giả thuyết H_0 ở mức $\alpha = 0,01$. Tra bảng F ở mức $\alpha = 0,01$ với độ tự do tử số và mẫu số lần lượt là 2 và 21 ta có giá trị $F_{\text{tra bảng}} = 5,78$. Ở mức $\alpha = 0,01$ giá trị $F_{\text{tính toán}} > \text{giá trị } F_{\text{tra bảng}}$, chúng ta kết luận các mức protein khác nhau trong khẩu phần có ảnh hưởng đến tăng trọng của lợn lai F_1 (Móng Cái $\times$ Đại Bạch) ở mức ý nghĩa $\alpha = 0,01$.

Ở phần kiểm tra giả thuyết và lực thống kê, chúng ta đã biết bước cuối cùng của quá trình kiểm tra giả thuyết thống kê là quyết định chấp nhận hay bác bỏ giả thuyết H_0 . Bước này có thể thực hiện theo ba cách. Ở đây, chúng tôi đề cập đến hai cách: Cách thứ nhất, từ giá trị α mà đã được lựa chọn ở bước thứ 2 của quá trình kiểm tra giả thuyết và độ tự do (df_B và df_W) chúng ta tra bảng để có giá trị F tra bảng, từ đó so sánh giá trị F tính toán và F tra bảng để đi đến kết luận chấp nhận hay bác bỏ giả thuyết H_0 . Đây là cách mà chúng ta vừa thực hiện ở trên. Cách thứ hai là từ giá trị F tính toán và độ tự do chúng ta tra bảng để có được giá trị xác suất (P). Chúng ta so sánh giá trị P vừa mới tra được với mức α lựa chọn ở bước thứ 2 của quá trình kiểm tra giả thuyết. Nếu giá trị P vừa tra được nhỏ hơn hoặc bằng giá trị α nêu ở bước 2 thì chúng ta bác bỏ giả thuyết H_0 . Ở ví dụ 4.3 (ví dụ về ảnh hưởng của các mức protein đến tăng trọng), với giá trị F tính toán 6,02 và độ tự do (2, 21), chúng ta có giá trị $P = 0,009$. Vì $0,009 < 0,005$ chúng ta bác bỏ giả thuyết H_0 . Cách thứ 2 này thông thường được các phần mềm xử lý thống kê sử dụng. Điều này có nghĩa nếu chúng ta xử lý thống kê bằng các phần mềm thì chúng ta không cần phải tra bảng F.

Ngoài ra còn một phương pháp khác có thể quyết định chấp nhận hay bác bỏ H_0 là dựa vào khoảng tin cậy của sự sai khác (Nội dung này sẽ được trình bày ở phần sau của chương này).

4.2.2.4. Trình bày kết quả phân tích phương sai

Kết quả ANOVA thường được trình bày trong một bảng tóm tắt. Các giá trị trong bảng tóm tắt có quan hệ với nhau. Kết quả ANOVA ở ví dụ 4.3 được trình bày ở bảng 4.3a. Một số phần mềm xử lý số liệu khác có thể có cách trình bày như ở bảng 4.3b.

Bảng 4.3a. Tóm tắt phân tích phương sai về ảnh hưởng của các mức protein khác nhau trong khẩu phần đến tăng trọng của lợn F1 (Móng Cái × Đại Bạch)

Nguồn biến động	SS	df	MS	F tính toán	Giá trị P	F tra bảng ở mức $\alpha = 0,05$
Giữa các nghiệm thức	114,4375	2	57,21875	6,02	0,009	3,4668
Trong các nghiệm thức	199,7188	21	9,510417			
Tổng số	314,1563	23				

Bảng 4.3b. Tóm tắt phân tích phương sai về ảnh hưởng của các mức protein khác nhau trong khẩu phần đến tăng trọng của lợn F1 (Móng Cái × Đại Bạch)

Nguồn biến động	SS	df	MS	F tính toán
Giữa các nghiệm thức	114,4375	2	57,21875	6,02*
Trong các nghiệm thức	199,7188	21	9,510417	
Tổng số	314,1563	23		

Ghi chú: * $P < 0,01$

Đa số các phần mềm xử lý số liệu đề cập rõ nguồn biến động. Ở ví dụ 4.3 nguồn biến động giữa các nghiệm thức được thay thế bằng “Mức protein” và nguồn biến động trong các nhóm được thay thế bằng “Sai số ngẫu nhiên”.

Dù phần mềm xử lý số liệu nào đi nữa thì một bảng tóm tắt phân tích phương sai thường thể hiện các nguồn biến động ở các hàng và thể hiện các giá trị tổng các bình phương, độ tự do, trung bình bình phương, giá trị F và mức độ tin cậy ở các cột.

Khi chỉ cần kiểm tra sự khác nhau giữa hai trị trung bình chúng ta có thể phân tích ANOVA hoặc kiểm tra t. Để tiến hành kiểm tra t, cần tính phương sai chung. Trong trường hợp phân tích phương sai, phương sai chung chính là giá trị MS_W . Hơn thế nữa để chấp nhận hay bác bỏ giả thuyết H_0 cần tính giá trị t tính toán hay F tính toán. Bình phương giá trị t tính toán chính là giá trị F tính toán. Trong trường hợp kiểm tra không khuynh hướng thì giá trị F tra bảng chính là bình phương giá trị t tra bảng. Để so sánh sự sai khác giữa hai giá trị trung bình, thông thường người ta dùng kiểm tra t đối với giả thuyết có khuynh hướng. Đối với giả thuyết không khuynh hướng chúng ta nên dùng phân tích ANOVA.

4.3. Đa so sánh

Kết quả phân tích ANOVA như ví dụ 4.3 (ví dụ về ảnh hưởng của các mức protein đến tăng trọng của lợn) ở trên đã trả lời câu hỏi nghiên cứu có hay không có ảnh hưởng của mức protein trong khẩu phần đến tăng trọng của lợn lai F1 (Móng Cái × Đại Bạch). Bác bỏ giả thuyết H_0 hay chấp nhận giả thuyết H_A có nghĩa là mức protein khác nhau

trong khẩu phần đã có ảnh hưởng đến khả năng tăng trọng. Tuy nhiên, kết quả phân tích ANOVA không chỉ ra được mức protein nào khác với mức protein nào về ảnh hưởng đến khả năng tăng trọng. Liệu cả 3 mức protein: 13%, 15% và 17% CP đều có ảnh hưởng khác nhau đến khả năng tăng trọng hay chỉ có mức protein 13% khác với mức 17% CP, còn mức protein 15% không khác với mức 17% CP hay liệu trung bình tăng trọng của lợn ở mức protein 13% và 15% khác với tăng trọng của lợn ở mức 17%CP. Để trả lời những câu hỏi này cần các phép phân tích tiếp theo, gọi là phân tích sau phân tích phương sai (*Post hoc*) và phép phân tích có kế hoạch. Khi tiến hành một thí nghiệm để trả lời các câu hỏi nghiên cứu, các nghiệm thức được lựa chọn và so sánh giữa các nghiệm thức sẽ trả lời câu hỏi nghiên cứu. Điều đó có nghĩa là nhiều khi chúng ta có kế hoạch so sánh giữa các nghiệm thức trước khi chúng ta tiến hành phân tích phương sai. So sánh trong trường hợp đó gọi là so sánh có kế hoạch.

Đa so sánh là một quá trình kiểm tra giả thuyết, do vậy nó bao gồm các bước của quá trình kiểm tra giả thuyết. Để tiến hành so sánh sự sai khác giữa các cặp nghiệm thức, ta cần có các giả thuyết H_0 và H_A bổ sung. Ví dụ, để so sánh ảnh hưởng của mức protein 13% và 17% CP đến khả năng tăng trọng ta có giả thuyết bổ sung ngoài giả thuyết của phân tích phương sai là: $H_0: \mu_{17\%CP} - \mu_{13\%CP} = 0$ và $H_A: \mu_{17\%CP} \neq \mu_{13\%CP}$.

4.3.1. Giả thuyết bổ sung và hệ số cân đối

Hệ số cân đối phụ thuộc vào giả thuyết bổ sung mà chúng ta muốn kiểm tra. Kết quả ANOVA ở ví dụ 4.3 cho thấy cần có các giả thuyết bổ sung để kiểm tra sự khác nhau giữa các trung bình với nhau. Cần lưu ý rằng nhiều khi chúng ta có kế hoạch kiểm tra các giả thuyết bổ sung trước khi có kết quả phân tích ANOVA. Với 3 nghiệm thức: 13%; 15% và 17% CP tương ứng với 3 giá trị μ_1 , μ_2 và μ_3 ta có thể có 6 giả thuyết bổ sung như sau:

1. So sánh giữa $\mu_{13\%}$ với $\mu_{15\%}$

$$H_0: \mu_{13\%} - \mu_{15\%} = 0$$

2. So sánh giữa $\mu_{13\%}$ với $\mu_{17\%}$

$$H_0: \mu_{13\%} - \mu_{17\%} = 0$$

3. So sánh giữa $\mu_{15\%}$ với $\mu_{17\%}$

$$H_0: \mu_{15\%} - \mu_{17\%} = 0$$

4. So sánh giữa kết hợp của $\mu_{13\%}$ với $\mu_{15\%}$ và $\mu_{17\%}$

$$H_0: \frac{\mu_1 + \mu_2}{2} - \mu_3 = 0$$

5. So sánh của kết hợp của $\mu_{13\%}$ với $\mu_{17\%}$ và $\mu_{15\%}$

$$H_0: \frac{\mu_1 + \mu_3}{2} - \mu_2 = 0$$

6. So sánh kết hợp của $\mu_{15\%}$ với $\mu_{17\%}$ và $\mu_{13\%}$

$$H_0: \frac{\mu_2 + \mu_3}{2} - \mu_1 = 0$$

Hệ số cân đối được phân chia cho các trung bình theo một cách có thể thể hiện được giả thuyết bổ sung cần kiểm tra và tổng các hệ số cân đối bằng 0. Nếu các nghiệm thức không tham gia vào các giả thuyết bổ sung cần kiểm tra thì hệ số cân đối của nghiệm thức đó bằng 0. Ví dụ, nếu muốn so sánh trung bình tăng trọng của nghiệm thức 13% và 15% CP, chúng ta có hệ số cân đối như sau:

$\mu_{13\%}$	$\mu_{15\%}$	$\mu_{17\%}$
1	- 1	0

Các hệ số cân đối được ký hiệu là w (weighted coefficient). Hệ số cân đối cho trung bình tăng trọng của lợn khi so sánh giữa mức protein 13% và 15% CP lần lượt là 1; -1 và 0. Đáp ứng được yêu cầu $w_{13\%} + w_{15\%} + w_{17\%} = 0$. Hệ số cân đối của trung bình tăng trọng ở mức 17% CP = 0, do tăng trọng của lợn ở mức protein này không tham gia vào phép so sánh giữa mức 13% và 15% CP.

Bảng 4.4. Hệ số cân đối của các phép so sánh tăng trọng của lợn được ăn 3 mức protein khác nhau (xem ví dụ 4.3)

Giả thuyết sau ANOVA	Hệ số cân đối		
	$\mu_{13\%}$	$\mu_{15\%}$	$\mu_{17\%}$
$H_0: \mu_1 - \mu_2 = 0$	1	- 1	0
$H_0: \mu_1 - \mu_3 = 0$	1	0	- 1
$H_0: \mu_2 - \mu_3 = 0$	0	1	- 1
$H_0: \frac{\mu_1 + \mu_2}{2} - \mu_3 = 0$	1/2	1/2	- 1
$H_0: \frac{\mu_1 + \mu_3}{2} - \mu_2 = 0$	1/2	- 1	1/2
$H_0: \frac{\mu_2 + \mu_3}{2} - \mu_1 = 0$	- 1	1/2	1/2

Chúng ta có thể chuyển các hệ số cân đối ở dạng số thập phân sang dạng số nguyên, miễn là đáp ứng yêu cầu $\sum w_i = 0$. Ví dụ, các hệ số cân đối mới cho phép so sánh giữa trung bình tăng trọng của lợn ở mức protein 13% và 15% CP với mức 17% CP lần lượt là 1; 1 và -2.

Quy tắc thiết lập hệ số cân đối nêu trên hoàn toàn có thể áp dụng cho trường hợp có nhiều hơn 3 trung bình cần so sánh. Ví dụ: kết quả phân tích ANOVA cho thấy có sự khác biệt về nồng độ amoniac phát xạ từ phân của lợn được ăn khẩu phần có 6 mức NSP khác nhau: 120; 140; 160; 180; 200 và 220 g/kg vật chất khô thức ăn. Chúng ta có thể có giả thuyết bổ sung để so sánh sự khác nhau về phát xạ amoniac từ phân của lợn được ăn 3 mức NSP thấp (120; 140 và 160 g/kg vật chất khô thức ăn) với 3 mức NSP cao (180; 200 và 220 g/kg vật chất khô thức ăn).

$$H_0: \frac{\mu_1 + \mu_2 + \mu_3}{3} - \frac{\mu_4 + \mu_5 + \mu_6}{3} = 0$$

Một tập hợp các hệ số có thể là:

μ_{120}	μ_{140}	μ_{160}	μ_{180}	μ_{200}	μ_{220}
1/3	1/3	1/3	- 1/3	- 1/3	- 1/3

Nếu ta nhân mỗi hệ số với 3 thì ta sẽ có:

μ_{120}	μ_{140}	μ_{160}	μ_{180}	μ_{200}	μ_{220}
1	1	1	- 1	- 1	- 1

4.3.2. So sánh có kế hoạch

Kết quả phân tích phương sai trình bày ở trên được giới hạn để kiểm tra có hay không có một sự khác nhau tổng thể giữa các trung bình của các nghiệm thức. Chúng ta thông thường quan tâm đến sự khác nhau giữa hai hay nhiều hơn các trung bình cụ thể nào đó. Trong trường hợp đó chúng ta dùng so sánh có kế hoạch. So sánh có kế hoạch hoàn toàn là một quá trình kiểm tra giả thuyết, do vậy nó bao gồm các bước cơ bản của một quá trình kiểm tra giả thuyết.

So sánh có kế hoạch bao gồm so sánh có kế hoạch trực giao/độc lập và so sánh có kế hoạch không trực giao. Bản chất của so sánh có kế hoạch độc lập là thay thế kiểm tra F tổng quát ($F = MS_B/MS_W$) bằng một tập hợp $t - 1$ các kiểm tra F đơn lẻ, với 1 độ tự do cho tử số và df_W độ tự do mẫu số.

Có hai trường hợp so sánh có kế hoạch nhưng không độc lập đó là: (1) Khi chúng ta muốn so sánh tất cả các nhóm nghiệm thức với một nhóm đối chứng; (2) Khi chúng ta muốn so sánh theo các cặp.

4.3.2.1. Phương sai, sai số chuẩn và khoảng tin cậy của hệ số tương phản

Hệ số so sánh hay còn gọi là hệ số tương phản (contrast), ký hiệu là C, giữa các trung bình nghiệm thức có thể được sử dụng để trả lời các câu hỏi nghiên cứu cụ thể. C được tính như sau:

$$C = \sum_{i=1}^t w_i \mu_i = w_1 \mu_1 + w_2 \mu_2 + \dots + w_t \mu_t \quad [4.17]$$

Các μ_i được tính từ các mẫu, cho nên ta có:

$$C = \sum_{i=1}^t w_i M_i = w_1 M_1 + w_2 M_2 + \dots + w_t M_t \quad [4.18]$$

Phương sai của hệ số so sánh ước tính $C = \sum_{i=1}^t w_i M_i$ được ước tính theo [4.19]

$$S_C^2 = S^2 \left[\sum_{i=1}^t \frac{w_i^2}{n_i} \right] = S^2 \left[\frac{w_1^2}{n_1} + \frac{w_2^2}{n_2} + \dots + \frac{w_t^2}{n_t} \right] \quad [4.19]$$

Nếu các n_i bằng nhau ta có:

$$S_C^2 = \frac{S^2}{n} [w_1^2 + w_2^2 + \dots + w_t^2] \quad [4.20]$$

Trong các công thức trên S^2 chính là giá trị MS_W hay MS_E từ phân tích phương sai.

Sai số chuẩn của hệ số tương phản chính là căn bậc hai của phương sai của nó và được tính theo công thức 4.21:

$$S_C = \sqrt{S_C^2} \quad [4.21]$$

Giả sử chúng ta đã có kế hoạch để so sánh ba trung bình tăng trọng khác nhau của lợn được ăn ba mức protein thô khác nhau trong khẩu phần (xem ví dụ 4.3): Trong ví dụ trên ta đã có trung bình tăng trọng của lợn tương ứng với ba mức protein là $M_1 = 6,875$; $M_2 = 11,125$ và $M_3 = 11,813$. Nếu chúng ta muốn so sánh tăng trọng của lợn ở mức protein 13% và 15% CP, ta có ba hệ số cân đối là:

$\mu_{13\%}$	$\mu_{15\%}$	$\mu_{17\%}$
1	- 1	0

Do vậy, $C_1 = (1)(\bar{X}_{13\%}) + (-1)(\bar{X}_{15\%}) + (0)(\bar{X}_{17\%}) = (1)(6,875) + (-1)(11,125) + (0)(11,813) = -4,25$.

Nếu chúng ta muốn so sánh tăng trọng của lợn ở mức protein 13% và 15% với mức 17% CP. Các hệ số cân đối là:

$\mu_{13\%}$	$\mu_{15\%}$	$\mu_{17\%}$
1	1	- 2

Do vậy, $C_2 = (1)(6,875) + (1)(11,125) + (-2)(11,813) = -5,626$

Phương sai của C_1 là:

$$S_{C_1}^2 = \frac{S^2}{n} [w_1^2 + w_2^2 + \dots + w_t^2] = \frac{9,510417}{8} [1^2 + (-1)^2 + 0^2] = 2,38$$

Phương sai của C_2 là:

$$S_{C_2}^2 = \frac{S^2}{n} [w_1^2 + w_2^2 + \dots + w_t^2] = \frac{9,510417}{8} [1^2 + (1)^2 + (-2)^2] = 7,13$$

Sai số chuẩn của C_1 và C_2 tương ứng là: 1,54 và 2,67.

Khoảng tin cậy của hệ số tương phản được tính theo công thức 4.22:

$$C \pm t_{\alpha/2, (N-t)}(S_C) \quad [4.22]$$

Khoảng tin cậy 95% của C_1 là: $-4,25 \pm t_{0,025, 21} 1,54 = -4,25 \pm 3,20$

Khoảng tin cậy 95% của C_2 là: $-5,626 \pm t_{0,025, 21} 2,67 = -5,626 \pm 5,55$

4.3.2.2. Tổng các bình phương của hệ số so sánh

Tổng các bình phương của hệ số so sánh của các nghiệm thức xác định lượng biến động của các quan trắc được giải thích bởi hệ số so sánh đó. Tổng các bình phương của một hệ số so sánh ($C = \sum_{i=1}^t w_i \mu_i$) được tính theo công thức 4.23:

$$SS_C = \frac{\left(\sum_{i=1}^t w_i M_i \right)^2}{\sum_{i=1}^t (w_i^2 / n_i)} \quad [4.23]$$

Nếu tất cả n_i bằng nhau ta sử dụng công thức 4.24:

$$SS_C = n \frac{\left(\sum_{i=1}^t w_i M_i \right)^2}{\sum_{i=1}^t w_i^2} \quad [4.24]$$

Độ tự do của tổng các bình phương của các hệ số so sánh = 1. Điều đó có nghĩa $SS_C = MS_C$.

4.3.2.3. Kiểm tra các hệ số so sánh bằng kiểm tra F và kiểm tra Student t

Để có thể kiểm tra các giả thuyết bổ sung như đã nêu trong phần 4.3.1. chúng ta có thể dùng kiểm tra F. Giá trị F tính toán được tính theo công thức 4.25 hay 4.26.

$$F_{\text{tính toán}} = \frac{MS_C}{MS_W} \quad [4.25]$$

Hay

$$F_{\text{tính toán}} = \frac{C^2}{MS_W \left(\frac{w_1^2}{n_1} + \frac{w_2^2}{n_2} + \dots + \frac{w_t^2}{n_t} \right)} \quad [4.26]$$

Trong đó MS_W được lấy từ ANOVA tổng thể của tập hợp số liệu.

Ở phần 4.3.2.1, chúng ta đã tính được C_1 và C_2 lần lượt là -4,25 và -5,626. Giá trị $F_{\text{tính toán}}$ của hệ số so sánh khi chúng ta muốn so sánh tăng trọng của lợn ở mức protein 13% và 15% CP là:

$$F_{\text{tính toán}} = \frac{(-4,25)^2}{9,510417 \left(\frac{1^2}{8} + \frac{(-1)^2}{8} + \frac{0^2}{8} \right)} = 7,60$$

Giá trị $F_{\text{tính toán}}$ của hệ số so sánh khi chúng ta muốn so sánh tăng trọng của lợn ở mức protein 13% + 15% với mức 17% CP:

$$F_{\text{tính toán}} = \frac{(-5,626)^2}{9,510417 \left(\frac{(1)^2}{8} + \frac{(1)^2}{8} + \frac{(-2)^2}{8} \right)} = 4,44$$

Trong so sánh có kế hoạch, số lượng các so sánh là nhỏ và sai sót loại I không phải là một vấn đề quan tâm lớn. Do vậy, chúng ta có thể sử dụng cùng mức α cho mỗi so sánh như đã dùng cho ANOVA tổng thể. Giả sử rằng chúng ta đã quyết định dùng $\alpha = 0,05$ trước khi tiến hành ANOVA. Để tìm giá trị $F_{\text{tra bảng}}$ chúng ta cần có df tử số và df mẫu số. Bởi vì mỗi so sánh chỉ có hai trung bình hoặc hai kết hợp của các trung bình, cho nên df tử số luôn là $2 - 1 = 1$, df mẫu số chính là df_W . Điều đó có nghĩa giá trị $F_{\text{tra bảng}}$ chính là $F(1; 21; 0,05) = 4,32$. Cả hai so sánh đều có $F_{\text{tính toán}} > 4,32$ có nghĩa là ta bác bỏ giả thuyết bổ sung H_0 trong cả hai trường hợp. Có nghĩa là tăng trọng của lợn được ăn khẩu phần có mức protein 13% có tăng trọng sai khác với lợn được ăn khẩu phần có mức protein 15% tại mức $\alpha = 0,05$. Trung bình tăng trọng của lợn ở mức protein 13% và 15%CP sai khác có ý nghĩa với tăng trọng của lợn ở mức 17% CP ở mức $\alpha = 0,05$.

Thay vì kiểm tra F chúng ta có thể sử dụng kiểm tra Student t. Mối quan hệ giữa phân bố F và phân bố Student t được thể hiện qua 4.27:

$$t^2 = F \quad [4.27]$$

Phân bố Student t có $N - t$ độ tự do, trong khi đó phân bố F có 1 là độ tự do tử số và $N - t$ là độ tự do mẫu số.

Giá trị t tính toán của hệ số so sánh được tính theo công thức 4.28:

$$t_{\text{tính toán}} = \frac{C}{S_C} \quad [4.28]$$

Trong đó, S_C chính là sai số chuẩn của hệ số so sánh và được tính theo công thức 4.21.

4.3.2.4. So sánh trực giao

Thuật ngữ trực giao (orthogonality) có gốc từ tiếng Hy Lạp. *Ortho* có nghĩa là thẳng, là vuông. *Gon* có nghĩa là góc cạnh. Như vậy, trực giao theo gốc từ Hy Lạp có nghĩa là góc vuông. Thuật ngữ trực giao được hiểu theo nghĩa rộng hơn là độc lập (so với cái khác), không dư thừa hay không trùng lặp. Do vậy, nếu hai so sánh là độc lập với nhau thì ta gọi là so sánh trực giao. Đối nghịch với so sánh trực giao còn được gọi là so sánh không trực giao.

Đặt giả sử chúng ta có 4 giá trị trung bình ($\bar{X}_1; \bar{X}_2; \bar{X}_3$; và $\bar{X}_4$) từ 4 nghiệm thức cần so sánh. Sự so sánh thứ nhất: giữa 2 giá trị trung bình $\bar{X}_1$ và $\bar{X}_2$ được gọi là trực giao với sự so sánh thứ 2: giữa 2 giá trị trung bình $\bar{X}_3$ và $\bar{X}_4$, bởi vì kết quả của phép so sánh thứ nhất không ảnh hưởng gì đến kết quả của phép so sánh thứ 2. Chúng ta có thể kiểm tra hai so sánh là trực giao hay không bằng cách nhân mỗi hệ số cân đối của trung bình trong so sánh 1 với hệ số cân đối tương ứng trong so sánh 2. Sau đó, cộng các tích đó lại. Nếu tổng của các tích là 0 thì hai phép so sánh đó là trực giao. Nếu tổng của các tích là khác 0 thì hai phép so sánh đó là không trực giao.

Bảng 4.5. Kiểm tra tính độc lập của so sánh $\bar{X}_1$ với $\bar{X}_3$ và $\bar{X}_2$ với $\bar{X}_3$

So sánh	Hệ số cân đối			
	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$
So sánh 1: $\bar{X}_1 - \bar{X}_3$	1	0	- 1	0
So sánh 2: $\bar{X}_2 - \bar{X}_3$	0	1	- 1	0
So sánh 1 x So sánh 2	1 x 0	0 x 1	(- 1) x (- 1)	0 x 0

Tổng của các tích so sánh 1 x so sánh 2 = 1.

Đặt giả sử nếu chúng ta muốn so sánh giữa trung bình của $\bar{X}_2$ và $\bar{X}_3$ với $\bar{X}_3$ và $\bar{X}_2$ với $\bar{X}_3$, cần kiểm tra xem hai phép so sánh đó có trực giao với nhau hay không? Chúng ta cần lập bảng tương tự như bảng 4.5.

Bảng 4.6. Kiểm tra tính độc lập của so sánh giữa $\frac{\bar{X}_2 + \bar{X}_3}{2} - \bar{X}_4$ và $\bar{X}_2 - \bar{X}_3$

So sánh	Hệ số cân đối			
	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$
$\frac{\bar{X}_2 + \bar{X}_3}{2} - \bar{X}_4$ (1)	0	1/2	1/2	-1
$\bar{X}_2 - \bar{X}_3$ (2)	0	1	-1	0
So sánh 1 $\times$ So sánh 2	0 $\times$ 0	(1/2) $\times$ 1	(1/2) $\times$ (-1)	(-1) $\times$ 0

Tổng của các tích so sánh 1 $\times$ so sánh 2 = 0.

Khi tổng của các tích so sánh 1 với so sánh 2 = 0 có nghĩa hai phép so sánh như vậy là trực giao. Số lượng so sánh độc lập tối đa có thể cho một tập hợp có t trung bình là $t - 1$. Khi chúng ta có 4 nghiệm thức cần so sánh ($t = 4$) thì sẽ có tối đa ba so sánh độc lập. Tuy nhiên, không phải chỉ có một tập hợp có thể của $t - 1$ so sánh độc lập. Khi số lượng so sánh là nhỏ thì so sánh độc lập được ưu ái hơn so với so sánh không độc lập, bởi vì nó cung cấp các thông tin không trùng lặp. Nếu hai so sánh là không độc lập thì tổng số thông tin đạt được là ít hơn so với so sánh hai mẫu độc lập. Vì lý do này, nên người ta thường khuyến cáo rằng các so sánh có kế hoạch nên là so sánh độc lập. Mặt khác, trong kiểm tra *Post hoc* số so sánh quan tâm có khả năng có thể nhiều hơn $t - 1$, nếu như vậy các so sánh rõ ràng không thể độc lập với nhau tất cả.

Một đặc tính rất quan trọng của các hệ số so sánh trực giao là tổng các bình phương của các hệ số so sánh trực giao chính là tổng các bình phương của nghiệm thức trong phân tích ANOVA. Cho nên trong so sánh trực giao có kế hoạch, thay vì kiểm tra F tổng thể (MS_B/MS_W) trong phân tích ANOVA, chúng ta tiến hành $t - 1$ kiểm tra F , với độ tự do của tử số = 1 và độ tự do của mẫu số chính là độ tự do của $df_W = N - t$.

4.3.3. So sánh không có kế hoạch và không độc lập (Kiểm tra Scheffé)

Khi chúng ta không có kế hoạch so sánh và các so sánh của chúng ta có thể không độc lập thì số so sánh có khả năng là rất lớn, bao gồm so sánh giữa các cặp nghiệm thức và sự kết hợp giữa các nghiệm thức. Do vậy sử dụng $\alpha = 0,05$ hay thậm chí $\alpha = 0,01$ (Bước 2 trong quá trình kiểm tra giả thuyết) là không hợp lý, bởi vì sai sót loại I có thể lớn đến mức không chấp nhận được. Mặt khác nếu chúng ta sử dụng $\alpha =$ (tỷ lệ sai sót loại I hợp lý/số kiểm tra) có thể dẫn đến giá trị α quá bé làm cho chúng ta hầu như không thể bác bỏ H_0 , khi H_0 là sai.

Kiểm tra Scheffé được thiết kế bởi Scheffé (1953) và có khả năng không chế được xác suất chấp nhận ít nhất một sai sót loại I trong tập hợp các so sánh. Do vậy khi sử

dụng kiểm tra Scheffé chúng ta có thể kiểm tra nhiều so sánh theo ý muốn, kể cả những so sánh không có kế hoạch. Tuy nhiên khi sử dụng loại kiểm tra này thì gây mất mát lớn về lực thống kê (xác suất bác bỏ một giả thuyết H_0 sai). Nói cách khác chúng ta chỉ bác bỏ H_0 khi chúng ta có giá trị F tính toán rất lớn.

Trong kiểm tra Scheffé chúng ta chọn một mức α , thông thường đó là mức mà chúng ta đã sử dụng trong ANOVA. Mức α mà chúng ta chọn chính là sai sót loại I cho tất cả tập hợp các các so sánh. Ví dụ nếu ta chọn $\alpha = 0,05$ trong kiểm tra Scheffé thì chúng ta biết rằng xác suất của ít nhất một sai sót loại I cho tất cả các so sánh có khả năng là 0,05.

Nếu F tổng thể của ANOVA là có ý nghĩa ở mức α thì kiểm tra Scheffé luôn luôn đưa đến sai khác có ý nghĩa của ít nhất một cặp so sánh nào đó với mức α . Mặt khác nếu F tổng thể không có ý nghĩa ở mức α thì không có căn cứ gì để áp dụng kiểm tra Scheffé, bởi vì sẽ không có so sánh nào có ý nghĩa.

Tiến trình kiểm tra Scheffé giống như so sánh độc lập có kế hoạch với 01 điểm khác biệt là giá trị F quyết định. Trong so sánh có kế hoạch độc lập giá trị quyết định chính là giá trị F tra bảng $F_{(1, N-t)}$, trong phép kiểm tra Scheffé giá trị F quyết định là $(t-1)F_{(t-1, N-t)}$. Giá trị quyết định trong kiểm tra Scheffé lớn hơn so với giá trị quyết định $F_{(1, N-t)}$, ngoại trừ khi $t = 1$. Giả thuyết H_0 trong kiểm tra Scheffé là $H_0: C = 0$. Giả thuyết H_0 bị bác bỏ khi:

$$|C| > S_C \sqrt{(t-1)F_{df_w}^{t-1} \alpha} \quad [4.29]$$

Tiếp tục biến đổi ta có:

$$C^2 > S_C^2 (t-1) F_{df_w}^{t-1} \alpha \quad [4.30]$$

$$\frac{C^2}{S_C^2} > (t-1) F_{df_w}^{t-1} \alpha \quad [4.31]$$

$$F_{\text{tính toán}} = \frac{C^2}{MS_w \left(\frac{w_1^2}{n_1} + \frac{w_2^2}{n_2} + \dots + \frac{w_t^2}{n_t} \right)} > (t-1) F_{df_w}^{t-1} \alpha \quad [4.32]$$

Khoảng tin cậy của tất cả các hệ số so sánh có khả năng là:

$$CI = C \pm S_C \sqrt{(t-1) F_{df_w}^{t-1} \alpha} \quad [4.33]$$

Quay trở lại ví dụ 4.3 như đã đề cập từ phần đầu chương. Kết quả phân tích ANOVA cho thấy mức protein trong khẩu phần có ảnh hưởng đến tăng trọng của lợn ở mức $\alpha = 0,05$. Giả sử chúng ta muốn so sánh giữa trung bình tăng trọng ở mức protein 13% và kết hợp của trung bình tăng trọng ở mức 15% và 17%CP. Chúng ta cũng muốn so sánh trung bình tăng trọng của lợn ở khẩu phần 13% và 15% CP. Hệ số cân đối của 2 phép so sánh này như sau:

	$\mu_{13\%}$	$\mu_{15\%}$	$\mu_{17\%}$
So sánh 1	2	-1	-1
So sánh 2	1	-1	0
So sánh 1 x so sánh 2	2 x 1	(-1) x (-1)	(0) x (-1)

Tổng chéo của các hệ số cân đối = 3, như vậy hai so sánh này không trực giao. Tuy nhiên, kiểm tra Scheffé không yêu cầu điều kiện trực giao. Giá trị $F_{\text{tính toán}}$ của so sánh 1 và so sánh 2 lần lượt là:

So sánh 1:

$$C = 2(6,875) + (-1)(11,125) + (-1)(11,813) = -9,188$$

$$F = \frac{(-9,188)^2}{9,510417\left(\frac{2^2}{8} + \frac{(-1)^2}{8} + \frac{(-1)^2}{8}\right)} = 11,835$$

So sánh 2 (như đã tính ở phần 4.3.2.3)

$$F_{\text{tính toán}} = \frac{(-4,25)^2}{9,510417\left(\frac{1^2}{8} + \frac{(-1)^2}{8} + \frac{0^2}{8}\right)} = 7,60$$

Tiếp theo chúng ta tính toán giá trị quyết định ở mức $\alpha = 0,05$. $F_{\text{quyết định}} = (3 - 1)F_{0,05}(2, 21) = 2 \times 3,4668 = 6,9336$. Như vậy, chúng ta bác bỏ giả thuyết H_0 ở cả hai so sánh. Chúng ta tiếp tục tính toán giá trị quyết định ở mức $\alpha = 0,01$. $F_{\text{quyết định}} = (3 - 1)F_{0,01}(2, 21) = 2 \times 5,7804 = 11,5608$. Như vậy, chỉ có so sánh thứ nhất có F lớn hơn 11,5608. Do vậy, sự khác nhau giữa tăng trọng của lợn ở mức 13%CP và sự kết hợp tăng trọng của lợn ở mức 15 và 17%CP ở mức $\alpha = 0,01$, nhưng sự khác nhau giữa tăng trọng của lợn được ăn khẩu phần 13% và 15%CP là không có ý nghĩa ở cùng mức $\alpha = 0,01$.

4.3.4. So sánh có kế hoạch và không độc lập (Kiểm tra Dunnett')

Trong thí nghiệm, nhiều khi một nghiệm thức được xem là đối chứng của tất cả các nghiệm thức còn lại. Do vậy, nhiều khi chúng ta muốn so sánh tất cả các nghiệm thức với đối chứng. Dunnett (1955) đã giới thiệu quy trình để thực hiện phép kiểm tra này.

Để có thể kiểm tra Dunnett (D) chúng ta tính giá trị D theo công thức 4.34.

$$D = d_{\alpha, t, df_w} \sqrt{\frac{2S^2}{n}} \quad [4.34]$$

Trong đó:

- d_{α, t, df_w} là giá trị tra bảng kiểm tra Dunnett
- n : số lần lặp lại
- S^2 chính là MS_W

Nếu số lần lặp lại không giống nhau, Hsu (1996) đề nghị sử dụng công thức 4.35.

$$D = d_{\alpha, t, df_w} \sqrt{2S^2 \left[\frac{1}{n_i} - \frac{1}{n_c} \right]} \quad [4.35]$$

Trong đó n_i và n_c là số lần lặp lại của nghiệm thức i và đối chứng.

Khoảng tin cậy hai phía của sự khác nhau giữa trung bình của mỗi nghiệm thức (μ_i) với trung bình của đối chứng (μ_c) được tính theo công thức 4.36.

$$M_i - M_c \pm D \quad [4.36]$$

Cận dưới của khoảng tin cậy nếu trung bình của nghiệm thức lớn hơn so với trung bình của đối chứng được tính theo công thức 4.37.

$$M_i - M_c - D \quad [4.37]$$

Cận trên của khoảng tin cậy nếu trung bình của nghiệm thức lớn hơn so với trung bình của đối chứng được tính theo công thức 4.38.

$$M_i - M_c + D \quad [4.38]$$

Nếu chúng ta muốn biết trung bình của một nghiệm thức có khác biệt với trung bình của đối chứng ở mức α hay không chúng ta có thể dùng kiểm tra 02 phía hoặc kiểm tra 01 phía.

Trường hợp kiểm tra hai phía: $H_0: \mu_i = \mu_c$; $H_A: \mu_i \neq \mu_c$. Giả thuyết H_0 sẽ bị bác bỏ khi:

$$|M_i - M_c| > D \quad [4.39]$$

Trường hợp kiểm tra một phía: $H_0: \mu_i \leq \mu_c$; $H_A: \mu_i > \mu_c$. Giả thuyết H_0 sẽ bị bác bỏ khi:

$$M_i - M_c > D \quad [4.40]$$

Trường hợp kiểm tra một phía: $H_0: \mu_i \geq \mu_c$; $H_A: \mu_i < \mu_c$. Giả thuyết H_0 sẽ bị bác bỏ khi:

$$M_i - M_c < -D \quad [4.41]$$

Ví dụ 4.4. Một thí nghiệm với 5 nghiệm thức (A, B, C, D, E) và 1 đối chứng (F) được tiến hành trên 24 đơn vị thí nghiệm. Số lần lặp lại/nghiệm thức hoặc đối chứng bằng 4. Giá trị trung bình của các nghiệm thức và đối chứng là: 8,29; 7,23; 7,54; 8,10; 8,59 và 7,10 lần lượt cho A, B, C, D, E và F. MS_W với độ tự do 18 là 0,08. Hãy so sánh sự sai khác giữa mỗi nghiệm thức và đối chứng ở mức $\alpha = 0,05$.

$$\text{Giá trị } D = d_{\alpha, t, df_w} \sqrt{\frac{2S^2}{n}} = d_{0,05; 5; 18} \sqrt{\frac{2 \times 0,08}{4}} = 2,76 \times 0,2 = 0,55$$

Nếu chúng ta so sánh sự sai khác của giữa trung bình của nghiệm thức và đối chứng với giá trị D thì có thể đưa ra kết luận có hay không có sự sai khác giữa trung bình nghiệm thức và đối chứng, tuy nhiên không nêu ra được mức độ của sự sai khác. Nếu chúng ta căn cứ vào khoảng tin cậy thì có thể đưa ra kết luận về sự sai khác giữa trung bình của nghiệm thức và đối chứng, đồng thời có thể kết luận về mức độ của sự sai khác.

Bảng 4.7. Kết quả kiểm tra Dunnett so sánh các nghiệm thức (A, B, C, D và E) với đối chứng F

Đối chứng/nghiệm thức	Trung bình	$M_i - M_c$	95%CI	$ M_i - M_c $	Khác với đối chứng*
F	7,10	-	-	-	-
A	8,29	1,19	(0,64; 1,74)	1,19	Có
B	7,23	0,13	(- 0,42; 0,68)	0,13	Không
C	7,54	0,44	(- 0,11; 0,99)	0,44	Không
D	8,10	1,00	(0,45; 1,55)	1,00	Có
E	8,59	1,49	(0,94; 2,04)	1,49	Có

* nếu $|M_i - M_c| > D = 0,55$ thì giá trị trung bình của nghiệm thức khác với đối chứng.

4.3.5. So sánh theo cặp của tất cả các nghiệm thức

4.3.5.1. Kiểm tra Tukey (HSD)

Trong nghiên cứu nhiều khi chúng ta muốn so sánh từng cặp nghiệm thức với nhau. Đặt giả sử một thí nghiệm có t nghiệm thức như vậy chúng ta sẽ có $t(t - 1)/2$ số so sánh giữa các cặp nghiệm thức $\mu_i - \mu_j$ với $i \neq j$. Một phương pháp so sánh sự khác nhau giữa các cặp nghiệm thức được Tukey phát triển vào 1949. Phương pháp này còn có tên gọi khác là sai khác trung thực có ý nghĩa, ký hiệu là HSD (honestly significant difference).

Kiểm tra Tukey yêu cầu số lần lặp lại (n) ở các nghiệm thức phải bằng nhau.

Kiểm tra Tukey bao gồm 3 bước cơ bản:

- Tính giá trị $q_{\text{tính toán}}$:

$$q_{\text{tính toán}} = \frac{M_i - M_j}{\sqrt{\frac{MS_w}{n}}} \quad [4.42]$$

Trong đó M_i là trung bình của nghiệm thức có giá trị trung bình lớn hơn, nói cách khác $M_i - M_j$ luôn luôn ≥ 0 .

- Tìm giá trị $q_{\text{tra bảng}}$:

Tra bảng sai khác Student hóa để tìm giá trị $q_{\alpha, t, dfw}$ với $P(q \geq q_{\alpha, t, dfw}) = \alpha$.

- So sánh giá trị $q_{\text{tính toán}}$ với giá trị $q_{\text{tra bảng}}$, nếu giá trị $q_{\text{tính toán}} > q_{\text{tra bảng}}$ ở mức α thì sự khác nhau giữa hai trung bình của hai nghiệm thức có ý nghĩa ở mức α , hay bác bỏ giả thuyết H_0 . Giả thuyết H_0 trong kiểm tra Tukey là $\mu_i - \mu_j = 0$ hay $\mu_i = \mu_j$ và giả thuyết H_A là $\mu_i \neq \mu_j$.

Với phương pháp như trên chúng ta phải luôn tính giá trị $q_{\text{tính toán}}$ cho các cặp so sánh khác nhau bởi vì với các cặp so sánh khác nhau thì tử số của công thức xác định $q_{\text{tính toán}}$ là khác nhau. Chúng ta có thể biến đổi công thức 4.42 để có một phương pháp kiểm tra Tukey đơn giản hơn.

Để sự sai khác giữa 2 giá trị trung bình của 2 nghiệm thức có ý nghĩa

$$\frac{M_i - M_j}{\sqrt{\frac{MS_w}{n}}} > q_{\text{tra bảng}} \text{ hay } M_i - M_j > q_{\text{tra bảng}} \sqrt{\frac{MS_w}{n}} \quad [4.43]$$

$$q_{\text{tra bảng}} \sqrt{\frac{MS_w}{n}} = \text{HSD} \quad [4.44]$$

Như vậy thủ tục kiểm tra Tukey bao gồm tính toán giá trị HSD. Sau đó so sánh giá trị chênh lệch giữa hai giá trị trung bình của hai nghiệm thức cần so sánh với giá trị HSD. Nếu chênh lệch giữa các trung bình $> \text{HSD}$, chúng ta bác bỏ giả thuyết H_0 , hay giá trị trung bình của hai nghiệm thức là sai khác ở mức α . Khi sự chênh lệch giữa hai giá trị trung bình của hai nghiệm thức $\leq \text{HSD}$ thì không có sự khác biệt giữa hai giá trị trung bình của 02 nghiệm thức ở mức α .

Cũng từ giá trị HSD và giá trị $M_i - M_j$ chúng ta có thể tính được khoảng tin cậy ở mức $100(1 - \alpha)\%$ cho giá trị sai khác giữa các cặp trung bình của các cặp nghiệm thức.

$$CI_{PC} = M_i - M_j \pm \text{HSD}_{(t, \alpha)} \text{ hay } M_i - M_j \pm q_{(\alpha, t, df_w)} \sqrt{\frac{MS_w}{n}} \text{ với } M_i \geq M_j \quad [4.45]$$

Trong đó CI_{PC} chính là khoảng tin cậy của chênh lệch giữa các cặp trung bình.

Thay vì điều kiện $M_i \geq M_j$, chúng ta có thể sử dụng giá trị tuyệt đối của sự chênh lệch giữa $M_i - M_j$. Công thức 4.45 có thể được viết ở dạng 4.46.

$$CI_{PC} = |M_i - M_j| \pm q_{(\alpha, t, df_w)} \sqrt{\frac{MS_w}{n}} \quad [4.46]$$

Chúng ta tiến hành kiểm tra Tukey để so sánh tăng trọng của lợn ở các mức protein khác nhau ở ví dụ 4.3. Để có thể tính được giá trị HSD ngoài các tham số đã biết như MS_w và n , chúng ta cần tra bảng giá trị q ở mức $\alpha = 0,05$ với độ tự do 3 và 21. Tra bảng chúng ta có giá trị xấp xỉ 3,58. Như vậy ta có giá trị $\text{HSD} = 3,903$

$$\text{HSD} = q_{\text{tra bảng}} \sqrt{\frac{MS_w}{n}} = 3,58 \times \sqrt{\frac{9,510417}{8}} = 3,903$$

Bảng 4.8. Kết quả kiểm tra so sánh cặp theo phương pháp Tukey HSD đối với ví dụ 4.3

So sánh		$ M_i - M_j $	95% CI _{PC}		P
			Cận dưới	Cận trên	
13%CP	15%CP	4,25*	0,35	8,15	0,031
13%CP	17%CP	4,87*	0,97	8,78	0,013
15%CP	17%CP	0,62	- 3,28	4,53	0,914

Có hai cách để diễn dịch sự khác nhau giữa các cặp trung bình: (1) căn cứ vào giá trị HSD; (2) căn cứ vào khoảng tin cậy của sự sai khác giữa hai giá trị trung bình. Cách thứ nhất chỉ trả lời được câu hỏi có hay không có sự sai khác giữa 2 giá trị trung bình. Cách thứ hai không những chỉ trả lời được câu hỏi có hay có sự sai khác mà còn chỉ ra mức độ sai khác bằng cách căn cứ vào giá trị cận dưới của khoảng tin cậy. Kết quả ở bảng 4.8 cho thấy sự sai khác về tăng trọng của lợn ở khẩu phần 13% so với 15% ít hơn so với giữa 13% và 17%. Giá trị cận dưới của 2 cặp so sánh lần lượt là 0,35 và 0,97.

Bởi vì số kiểm tra Tukey là nhỏ hơn so với kiểm tra Scheffé cho nên kiểm tra Tukey có sức mạnh hơn. Nói một cách khác, một sự sai khác thực sự tồn tại giữa các trung bình quần thể thì kiểm tra Tukey dễ phát hiện hơn so với kiểm tra Scheffé.

Tóm lại nếu tập hợp các so sánh chỉ bao gồm các cặp so sánh thì kiểm tra Tukey là một lựa chọn tốt hơn Scheffé. Tuy nhiên, nếu tập hợp các so sánh bao gồm cả so sánh theo cặp và không theo cặp thì kiểm tra Scheffé nên được dùng kể cả cho kiểm tra theo cặp.

4.3.5.2. Kiểm tra FISHER LSD

Giống như HSD, kiểm tra Fisher LSD (sai khác ít nhất có ý nghĩa) được dùng để so sánh các cặp trung bình với nhau. Một thuận lợi của LSD là không yêu cầu các mẫu có dung lượng mẫu bằng nhau. Một thuận lợi khác là nó có khả năng bác bỏ H_0 nhạy hơn so với kiểm tra HSD.

Giả thuyết H_0 là $\mu_i = \mu_j$ và giả thuyết H_A là $\mu_i \neq \mu_j$. Giả thuyết có thể được kiểm tra bằng thống kê phân bố Student:

$$t_0 = \frac{M_i - M_j}{\sqrt{S^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad \text{với } i \neq j \quad [4.47]$$

Với sai sót loại I = α , phương sai S^2 có độ tự do df_w thì giả thuyết H_0 bị bác bỏ khi giá trị tuyệt đối của $t_0 \geq t_{\alpha/2, df_w}$.

Phương pháp kiểm tra sai khác ít nhất có ý nghĩa (LSD) được phát triển từ công thức 4.47. Đối với tất cả các cặp giá trị trung bình của các cặp nghiệm thức sai khác ít nhất có ý nghĩa là:

$$LSD = t_{\alpha/2, df_w} \sqrt{S^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad [4.48]$$

Trong đó S^2 chính là MS_w . Giả thuyết H_0 bị bác bỏ khi $|M_i - M_j| \geq LSD$

Cũng từ giá trị LSD và giá trị $M_i - M_j$ chúng ta có thể tính được khoảng tin cậy ở mức $100(1 - \alpha)\%$ cho giá trị sai khác giữa các cặp trung bình của các cặp nghiệm thức.

$$CI_{PC} = M_i - M_j \pm t_{\alpha/2, df_w} \sqrt{S^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \text{ với } M_i \geq M_j \quad [4.49]$$

Trong đó CI_{PC} chính là khoảng tin cậy của chênh lệch giữa các cặp trung bình.

Thay vì điều kiện $M_i \geq M_j$, chúng ta có thể sử dụng giá trị tuyệt đối của sự chênh lệch giữa $M_i - M_j$. Công thức 4.49 có thể được viết ở dạng 4.50.

$$CI_{PC} = |M_i - M_j| \pm t_{\alpha/2, df_w} \sqrt{S^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad [4.50]$$

Chúng ta tiến hành kiểm tra LSD để so sánh tăng trọng của lợn ở các mức protein khác nhau ở ví dụ 4.3. Để có thể tính được giá trị LSD ngoài các tham số đã biết như MS_w và n , chúng ta cần tra bảng giá trị $t_{\alpha/2, df_w}$. Ở mức $\alpha = 0,05$ với độ tự do 21. Tra bảng Student chúng ta có giá trị là 2,0796. Như vậy ta có giá trị $LSD = 3,20$

Bảng 4.9. Kết quả kiểm tra so sánh cặp theo phương pháp Fisher LSD đối với ví dụ 4.3

So sánh		$ M_i - M_j $	95%CI		P
			Cận dưới	Cận trên	
13%CP	15%CP	4,25	1,03	7,47	,012
13%CP	17%CP	4,87	1,66	8,09	,005
15%CP	17%CP	0,62	- 3,84	2,59	,690

Cách diễn giải kết quả hoàn toàn tương tự như diễn giải kết quả ở bảng 4.8.

Tóm lại có nhiều kiểm tra đa so sánh khác nhau. Lựa chọn một phương pháp kiểm tra thích hợp là rất khó bởi vì mỗi kiểm tra đều có ưu và nhược điểm của nó. Tùy theo giả thuyết nghiên cứu, thiết kế nghiên cứu để lựa chọn kiểm tra đa so sánh phù hợp. Phương pháp kiểm tra Tukey có khả năng giảm tối đa việc đưa ra các quyết định sai. Xác định khoảng tin cậy có khả năng suy diễn thống kê mạnh nhất, có khả năng suy diễn khuynh hướng và mức độ sai khác của các so sánh.

4.4. Bài tập

1. Hãy bình luận kết quả phân tích phương sai sau?

Bảng 4.10. Kết quả phân tích phương sai ảnh hưởng của các loại khẩu phần đến khối lượng lợn lai F₁ (Móng Cái × Đại bạch)

Biến độc lập: Khối Lượng

Nguồn biến động	SS	df	MS	F	P
Khẩu phần	27,125	3	9,042	0,064	0,978
Ngẫu nhiên	2825,833	20	141,292		
Tổng	2852,958	23			

2. Hãy bình luận kết quả phân tích phương sai sau?

Bảng 4.11. Bảng tổng hợp phân tích phương sai ảnh hưởng của mức protein trong khẩu phần đến tăng trọng của lợn

Nguồn biến động	SS	df	MS	F tính toán	P
Mức protein	114,4375	2	57,21875	6,02	0,009
Sai số ngẫu nhiên	199,7188	21	9,510417		
Tổng số	314,1563	23			

3. Hãy kiểm tra giả thuyết H₀ “không có sự khác nhau về khối lượng sơ sinh giữa cừu Phan Rang, cừu Dorper và cừu lai F₁ (Dorper × Phan Rang) bằng cách sử dụng số liệu ở bảng 4.12. Nếu bác bỏ giả thuyết H₀, hãy sử dụng một số phân tích đa so sánh để kiểm tra sự sai khác về khối lượng sơ sinh của cừu Phan Rang, Dorper và F₁ (Dorper × Phan Rang)?

Bảng 4.12. Khối lượng sơ sinh của cừu Phan Rang, Dorper và F₁ (Dorper × Phan Rang)

Cừu Phan Rang	Cừu Dorper	Cừu lai F ₁	Cừu Phan Rang	Cừu Dorper	Cừu lai F ₁
1,9	2,9	3,10	2,1	3,4	2,70
2,2	3,4	2,80	2,2	3	3,00
2,3	3,1	2,90	2,3	2,8	2,70
2,0	3	3,10	2,0	3,2	1,90
2,0	3,4	2,60	2,1	3,2	3,20
2,0	2,6	1,90	2,4	3,2	1,90
2,1	3	1,80	2,1	3,3	1,80
2,2	2,9	2,70	2,4	3,4	3,10

Cừu Phan Rang	Cừu Dorper	Cừu lai F1	Cừu Phan Rang	Cừu Dorper	Cừu lai F1
2,0	3	2,20	2,0	3,4	2,10
2,1	3,3	3,10	2,0	2,4	2,00
2,3	3,2	3,30	2,5	2,6	2,90
2,4	2,9	2,90	2,1	2,8	3,10
2,1	3,2	2,10	2,6	3,3	2,90
2,3	3	2,20	2,1	3,1	2,10
2,0	3,1	2,40	2,0	2,9	2,90
2,4	2,90	3,10	2,5	2,60	2,90
2,3	3,30	2,10	1,9	2,80	2,90
2,5	3,30	1,90	2,2	3,20	2,00
2,4	3,10	2,60	2,0	3,10	2,90
2,6	3,30	3,20	2,5	3,00	1,80

Nguồn: Nguyễn Thành Vinh (2010).

- Một nghiên cứu đánh giá ảnh hưởng của các mức protein thô trong thức ăn tinh (10,1; 13,2; 16,9 và 20,2%) đến tăng trọng của bò Vàng Việt Nam trong giai đoạn vỗ béo. Thí nghiệm được tiến hành trên 24 con bò; 6 con/nghiệm thức. Giả định thí nghiệm được thiết kế theo kiểu CRD. Kết quả tăng trọng của bò được thể hiện ở bảng 4.13. Hãy phân tích ảnh hưởng của các mức protein thô trong thức ăn tinh đến tăng trọng của bò.

Bảng 4.13. Tăng trọng của bò Vàng Việt Nam khi được ăn các khẩu phần với mức protein trong thức ăn khác nhau

Nghiệm thức	Tăng trọng (g/ngày)	Nghiệm thức	Tăng trọng (g/ngày)
10,10%	331	16,90%	505
10,10%	649	16,90%	833
10,10%	547	16,90%	520
10,10%	651	16,90%	806
10,10%	662	16,90%	775
10,10%	655	16,90%	716
13,20%	691	20,20%	682
13,20%	829	20,20%	815
13,20%	502	20,20%	601
13,20%	739	20,20%	865
13,20%	667	20,20%	986
13,20%	818	20,20%	941

Nguồn: Đinh Văn Dũng (2010).

Chương 5

PHÂN TÍCH PHƯƠNG SAI HAI NHÂN TỐ

Nghiên cứu không chỉ dừng lại ở mức đánh giá ảnh hưởng của một nhân tố (ví dụ mức protein thô trong khẩu phần) đến biến phụ thuộc (ví dụ tăng trọng) mà có thể đồng thời nghiên cứu ảnh hưởng của hai hay nhiều hơn nhân tố cùng ảnh hưởng đến biến phụ thuộc.

Mức protein và năng lượng có ảnh hưởng như thế nào đến tăng trọng của lợn hay gia cầm? Các giống khác nhau và các mức protein khác nhau có ảnh hưởng như thế nào đến tăng trọng, khả năng chuyển hóa thức ăn của lợn? Các vùng sinh thái khác nhau và các giống bò khác nhau có ảnh hưởng như thế nào đến khả năng sinh trưởng và phát triển của bò cũng như khả năng sản xuất và hiệu quả sản xuất của nông hộ? Các câu hỏi như vậy cũng có nghĩa là một biến phụ thuộc có thể bị ảnh hưởng bởi hai biến độc lập hay còn gọi là nhân tố. Phân tích phương sai một nhân tố mà chúng ta đã thảo luận trong chương 4 được dùng để phân tích ảnh hưởng của một biến độc lập đến một hoặc một số biến phụ thuộc. Trong trường hợp phân tích ảnh hưởng của hai biến độc lập đến một hoặc một số biến phụ thuộc thì chúng ta tiến hành phân tích phương sai hai nhân tố. Chương này nhằm mục tiêu cung cấp kiến thức về lý do, cách tiến hành và cách diễn giải phân tích phương sai hai nhân tố. Ngoài ra, chương này cũng tập trung phân tích những điểm khác biệt giữa phân tích phương sai một nhân tố và hai nhân tố.

Về nguyên tắc, phân tích phương sai hai nhân tố giống với phân tích phương sai một nhân tố về lý do phân tích phương sai, các bước tiến hành cũng như cách trình bày kết quả phân tích. Điểm khác biệt cơ bản giữa phân tích phương sai hai nhân tố và một nhân tố là ước tính nguồn biến động giữa các nhân tố. Nguồn biến động này không chỉ do một nhân tố gây nên mà do cả hai nhân tố và tương tác giữa hai nhân tố gây nên. Điều này cũng có nghĩa phân tích phương sai hai nhân tố không chỉ có một cặp giá trị F tính toán và F tra bảng mà có ba cặp giá trị giữa F tính toán và F tra bảng.

5.1. Hai thí nghiệm một nhân tố hay một thí nghiệm hai nhân tố ?

Ví dụ 5.1. Một nhóm nghiên cứu quan tâm đến ảnh hưởng của hai mức protein khác nhau trong khẩu phần (17% và 15%) đến tăng trọng của hai giống lợn [F1 (Móng Cái $\times$ Đại Bạch) và Đại Bạch]. Cơ sở nghiên cứu có 10 con lợn Đại Bạch và 10 con lợn lai F1 (Móng Cái $\times$ Đại Bạch) và đầy đủ các trang thiết bị phục vụ cho triển khai nghiên cứu.

Một số thành viên của nhóm nghiên cứu đề xuất nghiên cứu ảnh hưởng của mức protein cao (17%) trong khẩu phần trên hai giống/tổ hợp lai lợn và sau đó tiếp tục nghiên cứu ảnh hưởng của mức protein thấp (15%) trên cùng hai giống lợn đó. Sau đó dùng kiểm tra t để so sánh tăng trọng của lợn ở hai giống khác nhau và hai mức protein khác nhau.

Mặc dầu tiến hành thí nghiệm như vậy cũng có thể trả lời được câu hỏi nghiên cứu nêu trên, nhưng nó lại có hai hạn chế: (1) thông thường ít hiệu quả hơn để tiến hành hai thí nghiệm so với một thí nghiệm mà cùng trả lời được câu hỏi nghiên cứu (xem thêm phần hiệu quả thiết kế thí nghiệm đã đề cập trong chương 4), (2) nếu kết quả chỉ ra rằng mức protein có ảnh hưởng khác nhau đến tăng trọng của lợn thuộc các giống khác nhau thì ta không có cách nào để lượng hóa được lượng tăng trọng mà do sự tương tác giữa mức protein và giống lợn gây nên, (3) tiến hành nhiều thí nghiệm riêng lẻ có thể gây tốn nguồn lực hơn so với tiến hành một thí nghiệm tổng hợp.

Một số thành viên khác trong nhóm nghiên cứu đề xuất sử dụng thiết kế thí nghiệm theo kiểu nhân tố. Kiểu thiết kế này có thể kiểm tra ảnh hưởng của hai biến độc lập cùng một lúc trong một thí nghiệm.

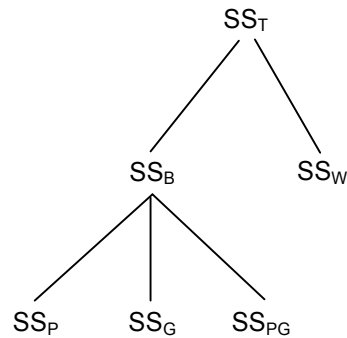
Chúng ta dùng ký hiệu P và G (hoặc có thể dùng ký hiệu khác) để ký hiệu cho hai biến độc lập trong thí nghiệm nêu trên, cụ thể P thay cho mức protein và G thay cho giống/tổ hợp lai lợn. Các con số nhỏ phía dưới ví dụ: 1, 2... thay cho các mức trong mỗi nhân tố. Cụ thể là có 4 kết hợp có thể giữa các mức của nhân tố G và P: P_1G_1 = Mức protein cao, tổ hợp lai F1; P_1G_2 = Mức protein cao, giống lợn Đại bạch; P_2G_1 = Mức protein thấp, tổ hợp lai F1; P_2G_2 = Mức protein thấp, giống lợn Đại bạch. Mỗi kết hợp xác định một nhóm các đơn vị thí nghiệm. Do vậy, có 4 nhóm trong nghiên cứu này hay còn gọi là 4 sự kết hợp giữa các nghiệm thức.

Một thí nghiệm nhân tố thường được xem xét dưới dạng số mức của mỗi nhân tố. Nếu ta ký hiệu a là số mức trong nhân tố P và b là số mức trong nhân tố G, lúc đó thí nghiệm ở ví dụ trên là 2×2 sự kết hợp các mức của nhân tố. Nếu có 3 mức protein thay vì hai mức như thí nghiệm trên thì thí nghiệm là 3×2 sự kết hợp các mức nhân tố.

5.2. Cơ sở của phân tích phương sai hai nhân tố

5.2.1. Các thành phần của phương sai

Giống như phân tích phương sai một nhân tố, tổng các bình phương tổng thể (SS_T) của các số liệu trong phân tích ANOVA hai nhân tố được phân chia làm hai thành phần cơ bản: (i) tổng các bình phương giữa các nhóm (SS_B), thể hiện sự biến động giữa các trung bình nhóm so với trung bình tổng thể và (ii) tổng các bình phương trong các nhóm (SS_W), thể hiện sự biến động của các giá trị trong các nhóm xung quanh các trung bình nhóm. Tuy nhiên, trong một thí nghiệm nhân tố nêu trên, biến động giữa các nhóm là một phần do ảnh hưởng của nhân tố P và một phần của nhân tố G. Do vậy SS_B được phân chia một cách cụ thể hơn nữa: tổng các bình phương nhân tố P (SS_P), tổng các bình phương nhân tố G (SS_G) và một tổng các bình phương cho tương tác giữa nhân tố P và nhân tố G (SS_{PG}). Sự phân chia SS_T thành các SS thành phần được thể hiện ở biểu đồ 5.1.



Biểu đồ 5.1. Sự phân chia tổng các bình phương trong ANOVA hai nhân tố

Tổng df cũng được phân chia theo cách tương tự như tổng bình phương. Giống như ANOVA một nhân tố, $df_T = N_T - 1$; $df_B = ab - 1$ và $df_W = N_T - ab$. Ở ví dụ 5.1 nêu trên có 5 cá thể (đơn vị thí nghiệm) trong một nhóm, có nghĩa là tổng số cá thể nghiên cứu (N_T) = 20, do đó các thành phần df như sau:

$$df_T = 20 - 1 = 19$$

$$df_B = 4 - 1 = 3$$

$$df_W = 20 - 4 = 16$$

df_B lại được phân chia một cách cụ thể hơn thành df_P , df_G và df_{PG} (tương tác giữa nhân tố P và nhân tố G).

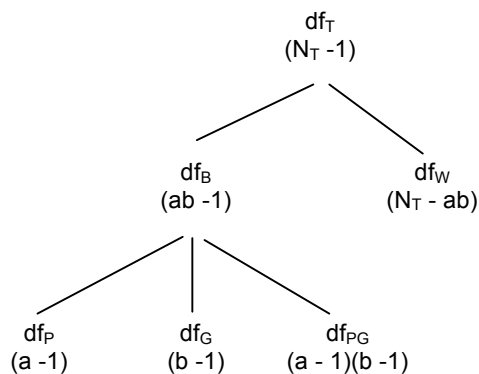
$$df_P = a - 1 = 2 - 1 = 1$$

$$df_G = b - 1 = 2 - 1 = 1$$

$$df_{PG} = (a - 1)(b - 1) = (1)(1) = 1$$

Như vậy: $df_P + df_G + df_{PG} = (a - 1) + (b - 1) + (a - 1)(b - 1) = ab - 1 = df_B$

Sự phân chia df_T được thể hiện ở biểu đồ 5.2.



Biểu đồ 5.2. Sự phân chia tổng độ tự do trong ANOVA hai nhân tố

5.2.2. Giả thuyết H_0 và H_A

Trong chương 5, ta đã biết phân tích ANOVA một nhân tố có một giả thuyết H_0 . Phân tích ANOVA hai nhân tố kiểm tra ba giả thuyết H_0 đồng thời.

Giả thuyết H_0 thứ nhất nêu rằng: Các trung bình quần thể dưới các mức của nhân tố P là bằng nhau. Giả thuyết H_A thứ nhất nêu rằng: Ít nhất một trung bình dưới các mức nhân tố P là khác với trung bình khác hay không phải tất cả các trung bình dưới các mức của nhân tố P là bằng nhau.

$$a. H_0: \mu_{P_1} = \mu_{P_2} = \dots = \mu_{P_a}$$

H_A : Ít nhất một trung bình dưới các mức nhân tố P là khác với trung bình khác = Không phải tất cả μ_P là bằng nhau.

Giả thuyết H_0 thứ hai nêu rằng: Các trung bình quần thể dưới các mức nhân tố G là bằng nhau. Giả thuyết H_A thứ hai nêu rằng: Ít nhất một trung bình dưới các mức nhân tố G là khác với trung bình khác hay không phải tất cả trung bình của nhân tố G là bằng nhau.

$$b. H_0: \mu_{G_1} = \mu_{G_2} = \dots = \mu_{G_b}$$

H_A : Ít nhất một trung bình dưới các mức nhân tố G là khác với trung bình khác = Không phải tất cả μ_G là bằng nhau.

Giả thuyết H_0 thứ ba nêu rằng: không có sự tương tác giữa nhân tố P và G. Giả thuyết H_A thứ ba nêu rằng có sự tồn tại của tương tác giữa nhân tố P và G.

$$c. H_0: \text{Không có ảnh hưởng của tương tác giữa nhân tố P và G.}$$

H_A : P và G có một ảnh hưởng tương tác giữa nhân tố P và G.

5.2.3. Kiểm tra giả thuyết H_0

Trong phân tích các thí nghiệm nhân tố, mỗi trung bình bình phương (MS) được tính cho mỗi nguồn ảnh hưởng. Như ANOVA một nhân tố, mỗi trung bình bình phương được tính bằng cách chia tổng các bình phương cho độ tự do tương ứng. Điều đó có nghĩa:

$$MS_P = \frac{SS_P}{df_P} \quad [5.1]$$

$$MS_G = \frac{SS_G}{df_G} \quad [5.2]$$

$$MS_{PG} = \frac{SS_{PG}}{df_{PG}} \quad [5.3]$$

Trung bình bình phương trong các nhóm được tính như ANOVA một nhân tố.

$$MS_W = \frac{SS_W}{df_W} \quad [5.4]$$

Nếu giả thuyết H_0 cho nhân tố P là đúng thì trung bình bình phương MS_P là tương đương với MS_W . Tuy nhiên, nếu H_0 là sai có nghĩa là nhân tố P thực sự có ảnh hưởng đến tình trạng nghiên cứu thì MS_P là lớn hơn MS_W . H_0 cho nhân tố P được kiểm tra bằng cách tính F như sau:

$$F_P = \frac{MS_P}{MS_W} \quad [5.5]$$

Xác suất (p) của F_P được tìm thấy bằng cách so sánh giá trị F_P tính toán với giá trị quyết định ở mức lựa chọn α và độ tự do (a - 1, $N_T - ab$). Nếu $p < \alpha$ thì ảnh hưởng của nhân tố P có ý nghĩa ở mức α và khi đó H_0 bị loại bỏ. Mặt khác nếu $p > \alpha$ thì ảnh hưởng của nhân tố P không có ý nghĩa ở mức α và H_0 không bị bác bỏ.

Lý luận hoàn toàn tương tự cho ảnh hưởng của nhân tố G và tương tác giữa P và G. Trong cả hai trường hợp, nếu H_0 là đúng thì trung bình bình phương của nhân tố nghiên cứu là tương đương với MS_W , nhưng nếu H_0 là sai thì trung bình bình phương của nhân tố nghiên cứu lớn hơn MS_W . Cụ thể đối với nhân tố G, nếu H_0 cho G là đúng, $MS_G = MS_W$ và nếu H_0 cho G là sai, $MS_G > MS_W$. Đối với tương tác giữa P và G, nếu H_0 cho tương tác là đúng, $MS_{PG} = MS_W$, nếu H_0 cho tương tác là sai, $MS_{PG} > MS_W$. Do vậy, kiểm tra mức độ ý nghĩa cho mỗi ảnh hưởng bao gồm tính các F sau:

$$F_G = \frac{MS_G}{MS_W} \quad [5.6]$$

$$F_{PG} = \frac{MS_{PG}}{MS_W} \quad [5.7]$$

Xác suất (p) của F_G được tìm thấy bằng cách so sánh F_G tính toán với giá trị quyết định ở mức lựa chọn α và độ tự do df (b - 1, $N_T - ab$). Xác suất của F_{PG} được tìm thấy bằng cách so sánh F_{PG} tính toán với giá trị quyết định ở mức lựa chọn α và df [(a - 1)(b - 1), $N_T - ab$]. Trong cả hai trường hợp nếu $p < \alpha$ thì ảnh hưởng có ý nghĩa ở mức α và H_0 bị bác bỏ và ngược lại.

5.3. Tiến hành phân tích phương sai hai nhân tố, một ví dụ thực tế

Giống như phân tích phương sai một nhân tố, phân tích phương sai hai nhân tố đòi hỏi các nhóm phải phân bố chuẩn và tương đồng về phương sai giữa các nhóm thuộc các nhóm ($\sigma^2_{A_1B_1} = \sigma^2_{A_1B_2} = \dots = \sigma^2_{A_aB_b}$). Khi số lượng mẫu lớn thì vẫn có thể phân tích ANOVA mặc dù hai điều kiện trên có thể không được đáp ứng.

Có rất nhiều phần mềm thống kê khác nhau có thể tiến hành phân tích phương sai hai nhân tố. Tuy nhiên, hiểu bản chất cũng như diễn dịch kết quả phân tích phương sai hai nhân tố là điều quan trọng. Tiến hành phân tích phương sai hai nhân tố không có sự trợ giúp của máy tính là cách tốt nhất để hiểu bản chất của phân tích phương sai hai nhân tố.

Các số liệu của thí nghiệm $a \times b$ nhân tố có thể được biểu diễn trong một bảng hai chiều (Bảng 5.1). Mỗi hàng của bảng thể hiện các số liệu (Giá trị tính trạng đo lường trên các đơn vị thí nghiệm) của mỗi mức của nhân tố P và mỗi cột thể hiện các số liệu của mỗi mức của nhân tố G. Chúng ta có thể biểu diễn số liệu theo chiều ngược lại. Các loại tổng cần thiết cho quá trình phân tích ANOVA cũng được thể hiện ở bảng 5.1. Các tổng này bao gồm tổng của các giá trị trong mỗi ô (sự kết hợp giữa các nghiệm thức), tổng của các giá trị mỗi hàng và tổng của các giá trị mỗi cột. Chúng ta ký hiệu ΣPG là tổng của các giá trị trong mỗi ô và dùng các chữ số phía dưới để thể hiện mức độ của nhân tố A và nhân tố B. Ví dụ, tổng các giá trị trong nhóm P_3G_5 là ΣP_3G_5 hay chính là tổng các giá trị của tính trạng của các đơn vị thí nghiệm nhận tác động của nghiệm thức P_3 và G_5 . Tổng các giá trị trong bất cứ hàng nào, ΣP là tổng giá trị của tính trạng xác định trên các đơn vị thí nghiệm nhận ảnh hưởng của mức nhân tố P trên tất cả mức độ của nhân tố G. Ví dụ ΣP_4 là tổng của các giá trị của các đơn vị thí nghiệm nhận ảnh hưởng của nghiệm thức P_4 trên tất cả mức của nhân tố G ($\Sigma P_4 = \Sigma P_4G_1 + \Sigma P_4G_2 + \dots + \Sigma P_4G_b$). Tương tự như vậy tổng của các giá trị trong bất cứ cột nào, ΣG là tổng của các giá trị của mức G với mọi mức của nhân tố P, ví dụ $\Sigma G_3 = \Sigma P_1G_3 + \Sigma P_2G_3 + \dots + \Sigma P_aG_3$.

Ví dụ 5.2. Một nhóm nghiên cứu quan tâm đến ảnh hưởng của việc tiêm hormon và giới tính đến hàm lượng canxi trong máu của gà thuộc hai giới tính khác. Thí nghiệm có hai nhân tố. Nhân tố xử lý hormon (Có, Không), nhân tố giới tính (Trống, Mái). Thí nghiệm được bố trí theo nhân tố giao nhau, 2×2 nhân tố. Mỗi sự kết giữa các mức của nhân tố có 5 lần lặp lại.

Hai nhân tố được gọi là giao nhau (cross) khi mỗi mức của một nhân tố được tìm thấy trong sự kết hợp với mỗi mức của nhân tố kia³. Mỗi sự kết hợp có 5 cá thể, như vậy ta có $N_T = 2 \times 2 \times 5 = 20$. Số lần lặp lại của các nhóm bằng nhau ($n = 5$), trong trường hợp như vậy người ta còn gọi là thiết kế thí nghiệm cân bằng (balanced experimental design). Đôi khi số lần lặp lại không bằng nhau trong mỗi sự kết hợp giữa các nhân tố, khi đó ta cần phải có một quy trình xử lý khác.

³ Chúng ta cần phân biệt sự khác nhau giữa bố trí thí nghiệm giao nhau (Cross Arrangement) và bố trí thí nghiệm phân tổ (Nested Arrangement).

Bảng 5.1. Ma trận số liệu của thí nghiệm nhân tố có a × b sự kết hợp giữa các mức của nhân tố

	Mức nhân tố G				Tổng các quan sát	Số quan sát
	G ₁	G ₂	...	G _b		
Mức nhân tố P	P ₁	Các giá trị của n đơn vị thí nghiệm trong nhóm P ₁ G ₁ Tổng = ΣP ₁ G ₁	Các giá trị của n đơn vị thí nghiệm trong nhóm P ₁ G ₂ Tổng = ΣP ₁ G ₂	Các giá trị của n đơn vị thí nghiệm trong nhóm P ₁ G _b Tổng = ΣP ₁ G _b	ΣP ₁	bn*
	P ₂	Các giá trị của n đơn vị thí nghiệm trong nhóm P ₂ G ₁ Tổng = ΣP ₂ G ₁	Các giá trị của n cá thể trong nhóm P ₂ G ₂ Tổng = ΣP ₂ G ₂	Các giá trị của n đơn vị thí nghiệm trong nhóm P ₂ G _b Tổng = ΣP ₂ G _b	ΣP ₂	bn
	...					
	P _a	Các giá trị của n đơn vị thí nghiệm trong nhóm P _a G ₁ Tổng = ΣP _a G ₁	Các giá trị của n đơn vị thí nghiệm trong nhóm P _a G ₂ Tổng = ΣP _a G ₂	Các giá trị của n đơn vị thí nghiệm trong nhóm P _a G _b Tổng = ΣP _a G _b	ΣP _a	bn
Tổng các quan sát	ΣG ₁	ΣG ₂		ΣG _b	ΣT	
Số các quan sát	an	an		an	N _T ** =	abn

Chú ý: * = n = Số lần lặp lại trong mỗi sự kết hợp của các nhân tố

** = N_T là tổng số các quan sát. Do có a × b sự kết hợp giữa các nghiệm thức và mỗi sự kết hợp có n quan sát nên N_T = a × b × n

Một điều cần lưu ý là khi phân tích phương sai hai nhân tố bằng một số phần mềm thống kê thì không nên có số lần lặp lại chênh lệch quá lớn giữa sự kết hợp của các nghiệm thức, bởi vì trong quá trình xử lý thì một số phần mềm gán cho các giá trị thiếu (các mẫu không có giá trị) một giá trị bằng trung bình của các giá trị còn lại trong nhóm. Khi đó thí nghiệm được hiểu là một thí nghiệm cân bằng, điều này là không thực tế. Sai sót sẽ lớn khi số mẫu càng không cân bằng. Ở một số phần mềm thống kê như Statitics thì nếu số mẫu chênh lệch nhau từ 50 trở lên thì chương trình không thể chạy. Tuy nhiên, đối với SPSS hay SAS thì dù có chênh lệch bao nhiêu đi nữa nó vẫn chạy được chương trình. Để khắc phục hiện tượng này, cần sử dụng các thuật toán và trình ứng dụng phù hợp trong các phần mềm thống kê để phân tích số liệu của các thí nghiệm không cân bằng.

Các tổng bình phương được tính như sau:

- *Tổng bình phương tổng thể*

$$SS_T = \sum X_T^2 - \frac{(\sum X_T)^2}{N_T} \quad [5.8]$$

$$SS_T = 11354,3 - \frac{(436,5)^2}{20} = 1827,7 \quad [5.9]$$

- *Tổng các bình phương giữa các nhóm*

$$\frac{(\sum A_1 B_1)^2}{n} + \frac{(\sum A_1 B_2)^2}{n} + \frac{(\sum A_2 B_1)^2}{n} + \frac{(\sum A_2 B_2)^2}{n} - \frac{(\sum X_7)^2}{N_7} \quad [5.10]$$

$$\frac{(74,4)^2}{5} + \frac{(60,6)^2}{5} + \frac{(162,6)^2}{5} + \frac{(138,9)^2}{5} - \frac{(436,5)^2}{20}$$

- *Tổng các bình phương trong các nhóm*

Chúng ta có thể tính SS_W một cách trực tiếp từ số liệu bằng cách tính tổng các bình phương trong các nhóm sau đó cộng lại. Trong tập hợp số liệu trên ta có 4 tổng bình phương trong các nhóm lần lượt là (24,86; 17,91; 278,87 và 44,73). Tổng của chúng sẽ là 366,3. Hoặc chúng ta có thể tính bằng công thức: Tổng các bình phương trong các nhóm = tổng các bình phương tổng thể - tổng các bình phương giữa các nhóm. Chúng ta cũng sẽ có kết quả tương tự: $SS_W = 1827,7 - 1461,3 = 366,3$.

Như đã đề cập ở trên, tổng biến động do nhân tố thí nghiệm tạo nên (tổng biến động giữa các nhóm) gồm có ba thành phần: hai nhân tố chính và tương tác giữa hai nhân tố. SS_A , SS_B , SS_{AB} .

- *Tổng các bình phương do nhân tố hormon (SS_A):*

$$SS_A = \frac{(\sum A_1)^2}{bn} + \frac{(\sum A_2)^2}{bn} - \frac{(\sum X_T)^2}{N_T} = \frac{(135,0)^2}{10} + \frac{(301,5)^2}{10} - \frac{(436,5)^2}{20} = 1386,1 \quad [5.11]$$

Bảng 5.2. Ma trận các số liệu trong thí nghiệm 2×2 nhân tố trong ví dụ về hàm lượng canxi trong máu gà dưới sự tác động của nhân tố hormon và giới tính

	Mái (B1)	Trống (B2)	Tổng
Không hormon (A1)	16,5	14,5	
	18,4	11,0	
	12,7	10,8	
	14,0	14,3	$\sum A_1 = 135,0$
	12,8	10,0	$bn = 10$
	$\sum A_1 B_1 = 74,4$	$\sum A_1 B_2 = 60,6$	
	$n = 5$	$n = 5$	
Có hormon (A2)	39,1	32,0	
	26,2	23,8	
	21,3	28,8	
	35,8	25,0	$\sum A_2 = 301,5$
	40,2	29,3	$bn = 10$
	$\sum A_2 B_1 = 162,6$	$\sum A_2 B_2 = 138,9$	

n = 5	n = 5	
$\Sigma B_1 = 237,0$	$\Sigma B_2 = 199,5$	$\Sigma X_T = 436,5$
an = 10	an = 10	$N_T = 20$

- Tổng các bình phương do nhân tố giới tính (SS_B):

$$SS_B = \frac{(\Sigma B_1)^2}{an} + \frac{(\Sigma B_2)^2}{an} - \frac{(\Sigma X_T)^2}{N_T} = \frac{(237,0)^2}{10} + \frac{(199,5)^2}{10} - \frac{(436,5)^2}{20} = 70,3 \quad [5.12]$$

- Tổng các bình phương do tương tác giữa nhân tố hormon và giới tính:

$$SS_{AB} = SS \text{ giữa các nhóm} - SS_A \text{ (nhân tố hormon)} - SS_B \text{ (nhân tố giới tính)} \\ = 1461,3 - 1386,1 - 70,3 = 4,9.$$

Sau khi tính được tổng các bình phương, bước tiếp theo là tính trung bình bình phương

- Trung bình bình phương của nhân tố hormon

$$MS_A = \frac{SS_A}{df_A} \text{ trong đó } df_A = a - 1 \quad [5.13]$$

- Trung bình bình phương của nhân tố giới tính

$$MS_B = \frac{SS_B}{df_B} \text{ trong đó } df_B = b - 1 \quad [5.14]$$

- Trung bình bình phương do tương tác giữa hormon và giới tính

$$MS_{AB} = \frac{SS_{AB}}{df_{AB}} \text{ trong đó } df_{AB} = (a - 1)(b - 1) \quad [5.15]$$

- Trung bình bình phương do ngẫu nhiên

$$MS_W = \frac{SS_W}{df_W} \text{ trong đó } df_W = N_T - ab \quad [5.16]$$

Thay giá trị vào các công thức nêu trên ta có giá trị của MS_A , MS_B , MS_{AB} , MS_W lần lượt là: 1386,1; 70,3; 4,9 và 22,89.

Bước tiếp theo tính các tỷ lệ F bằng cách chia các trung bình bình phương cho MS_W .

$$F_A = \frac{MS_A}{MS_W} = 60,5$$

Ta có $F_{0,05;1;16} = 4,49$ (giá trị quyết định của F ở mức $\alpha = 0,05$ và độ tự do 1 và 16). Như vậy $F_A \text{ tính toán} > \text{giá trị } F_{\text{quyết định}}$ cho nên ta bác bỏ giả thuyết H_0 có nghĩa là nhân tố A đã có ảnh hưởng (việc xử lý hormon đã có ảnh hưởng đến hàm lượng canxi máu) với $P < 0,05$ (Nếu ta tính bằng các phần mềm thống kê ví dụ PQRS thì $P = 0,00000080$).

$$F_B = \frac{MS_B}{MS_W} = 3,07$$

Ta có $F_{0,05;1;16} = 4,49$. Như vậy F_B tính toán < giá trị $F_{\text{quyết định}}$ cho nên ta không bác bỏ giả thuyết H_0 (giới tính không ảnh hưởng đến hàm lượng canxi máu của gà) với $0,05 < P < 0,1$ ($P = 0,099$).

$$F_{AB} = \frac{MS_{AB}}{MS_W} = 0,214$$

Ta có $F_{0,05;1;16} = 4,49$. Như vậy $F_{AB} < \text{giá trị } F_{\text{quyết định}}$ nên ta không bác bỏ giả thuyết H_0 (không có sự tương tác giữa hormon và giới tính lên hàm lượng canxi máu với $P > 0,05$ ($P = 0,65$)).

Kết quả phân tích phương sai hai nhân tố được trình bày trên một bảng tóm tắt ANOVA như sau (đa số các phần mềm thống kê hiện tại đều có cách trình bày giống như kết quả ở bảng 5.3).

Bảng 5.3. Tóm tắt ANOVA hai nhân tố

Nguồn biến động	SS	df	MS	F	P
Nhân tố A (hormon)	1386,1	1	1386,1	60,5	0,00000080
Nhân tố B (giới tính)	70,3	1	70,3	3,07	0,099
A x B	4,9	1	4,9	0,214	0,65
Ngẫu nhiên (trong các nhóm)	366,4	16	22,89		
Tổng	1827,69	19			

Qua bảng 5.3 hay kết quả phân tích phương sai hai nhân tố cho chúng ta biết yếu tố nào có ảnh hưởng có ý nghĩa và yếu tố nào không có ý nghĩa, nhưng nó lại không nói cho chúng ta biết khuynh hướng của sự sai khác. Để biểu diễn khuynh hướng của sự sai khác cần biểu diễn các giá trị trung bình (Bảng 5.4) trên các đồ thị hai chiều (Sơ đồ 5.3).

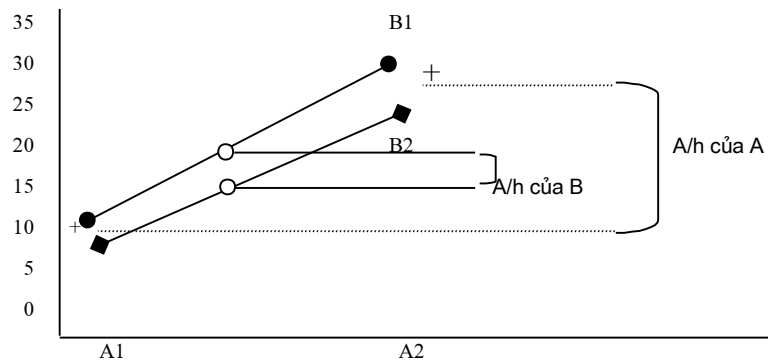
Bảng 5.4. Trung bình ô, cột, hàng của các số liệu trong ví dụ về ảnh hưởng của hormon và giới tính đến hàm lượng canxi máu của gà (mg/100ml)

	Mái (B_1)	Trống (B_2)	
Không hormon (A_1)	14,9	12,1	13,5
Có hormon (A_2)	32,5	27,8	30,2
	23,7	20,0	

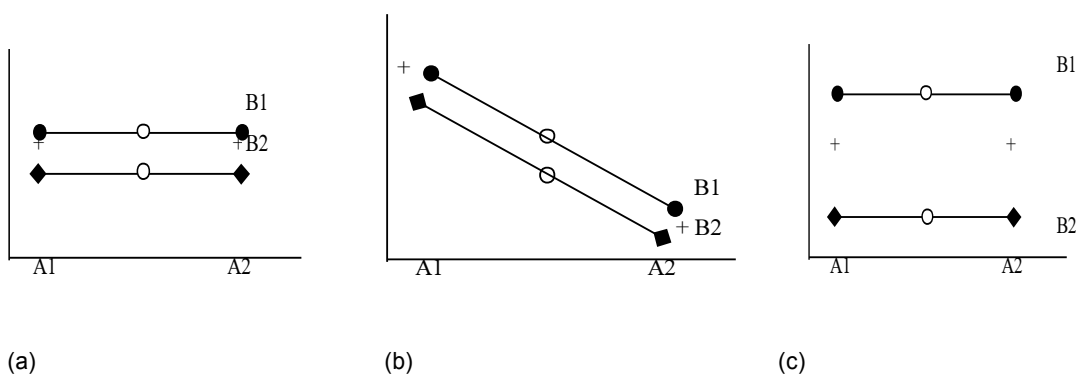
Ở trên sơ đồ 5.3a, hai mức độ A_1 và A_2 của nhân tố A (hormon) được thể hiện trên trục hoành và hai mức độ B_1 và B_2 (giới tính) được thể hiện trên trục tung. Hai trung

biên độ của B_1 (14,9 và 32,5) được thể hiện bằng hai chấm đen tròn và được nối với nhau bằng một đường thẳng. Hai trung bình độ của B_2 (12,1 và 27,8) được thể hiện bằng hai chấm đen hình thoi và được nối với nhau bằng một đường thẳng. Trung bình của tất cả số liệu trong mỗi mức độ của nhân tố A được thể hiện bằng dấu cộng và của nhân tố B được thể hiện bằng ô tròn. Khi đó ảnh hưởng của nhân tố A chính là khoảng cách thẳng đứng giữa hai dấu cộng và ảnh hưởng của nhân tố B chính là khoảng cách giữa hai ô tròn và sự không song song của các đường nói lên sự tương tác giữa nhân tố A và nhân tố B.

Kết quả phân tích ANOVA được nhìn thấy một cách rất rõ ràng hơn khi biểu diễn trên biểu đồ 5.3. Cụ thể là có một ảnh hưởng rất lớn của nhân tố A, một ảnh hưởng nhỏ của nhân tố B và một phần rất nhỏ của tương tác giữa A và B, được thể hiện bởi sự không song song của hai đường thẳng. Các trường hợp ảnh hưởng khác có thể xảy ra được thể hiện trên biểu đồ 5.3b.



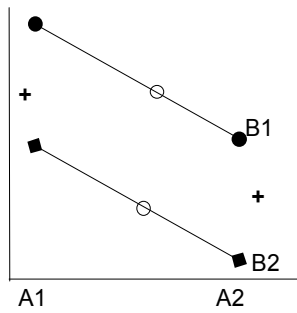
Biểu đồ 5.3a. Mô hình mô phỏng các giá trị trung bình nghiệm thức trong ANOVA hai nhân tố



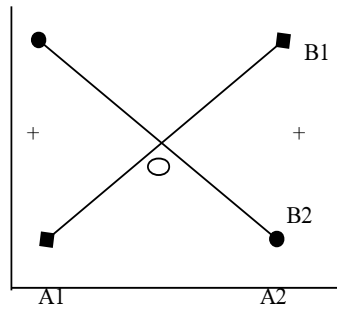
(a) Không có ảnh hưởng của A, một phần nhỏ ảnh hưởng của B (nếu không có ảnh hưởng của B hai đường sẽ trùng nhau) và không có tương tác giữa A và B

(b) Có ảnh hưởng lớn của A, ảnh hưởng nhỏ của B và không có tương tác giữa A và B

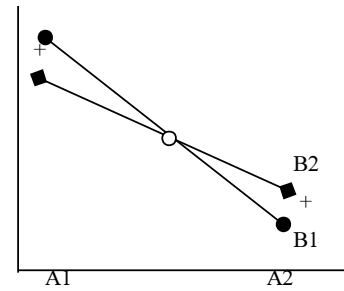
(c) Không có ảnh hưởng của A, có ảnh hưởng rất lớn của B và không có tương tác giữa A và B



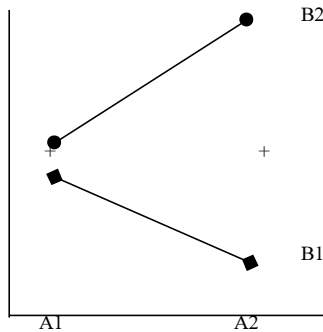
(d)
Có ảnh hưởng lớn của A và B,
không có tương tác giữa A và B



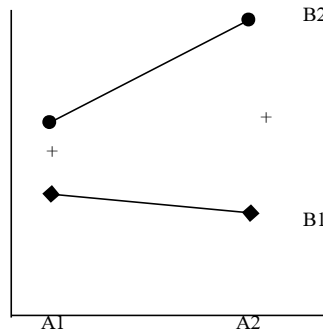
(e)
Không có ảnh hưởng của A và B
nhưng lại có tương tác giữa A và B



(f)
Có ảnh hưởng lớn của A, không
có ảnh hưởng của B và tương
tác nhỏ giữa A và B



(g)
Không có ảnh hưởng của A, ảnh
hưởng lớn của B và tương tác
chặt chẽ giữa A và B



(h)
Có ảnh hưởng của A, ảnh hưởng
lớn của B và tương tác lớn giữa
A và B

**Biểu đồ 5.3b. Mô hình mô
phỏng các giá trị trung bình
của các nghiệm thức trong
ANOVA hai nhân tố và mối
tương tác giữa chúng**

5.4. Các mô hình phân tích phương sai hai nhân tố

• Mô hình I

Phân tích phương sai hai nhân tố mà ở đó tất cả các nhân tố là cố định thì được gọi là mô hình ANOVA I. Trong những trường hợp như thế này, H_0 về không khác nhau giữa các mức của một nhân tố là được kiểm tra bằng cách sử dụng $F = MS_{nhân\ tố} / MS_{ngẫu\ nhiên}$. Thêm vào đó chúng ta có thể kiểm tra mức độ ý nghĩa của sự tương tác giữa các nhân tố trong mô hình I bằng cách dùng $F = MS_{tương\ tác} / MS_{ngẫu\ nhiên}$.

Trong phân tích phương sai hai nhân tố, các ảnh hưởng của một hoặc hai nhân tố là có ý nghĩa, nhưng ảnh hưởng của tương tác có thể có ý nghĩa hoặc có thể không có ý nghĩa. Đôi khi chúng ta gặp phải trường hợp mà ở đó sự tương tác có ý nghĩa xảy ra nhưng ảnh hưởng của mỗi yếu tố chính lại không có ý nghĩa.

• Mô hình II

Nếu một bố trí thí nghiệm nhân tố mà các mức của nhân tố là ngẫu nhiên thì ta gọi là mô hình ANOVA II (đây là một trường hợp khá hiếm). Trong trường hợp như vậy,

kiểm tra thống kê thích hợp cho ảnh hưởng của nhân tố là $F = MS \text{ nhân tố} / MS \text{ tương tác}$ (Bảng 5.4). Chúng ta kiểm tra sự tương tác như đã làm trước đây $F = MS \text{ tương tác} / MS \text{ ngẫu nhiên}$.

- Mô hình III

Nếu một bố trí thí nghiệm nhân tố có cả nhân tố cố định và nhân tố ngẫu nhiên, khi đó chúng ta nói rằng bố trí thí nghiệm thuộc loại *mô hình hỗn hợp* hay còn gọi là *ANOVA mô hình III*⁴. Các kiểm tra F thích hợp được thể hiện ở bảng 5.5.

Bảng 5.5. Giá trị $F_{\text{tính toán}}$ cho việc kiểm tra mức độ ý nghĩa của các nhân tố trong ANOVA hai nhân tố có lặp lại

Ảnh hưởng	Mô hình I (cả A và B đều cố định)	Mô hình II (cả A và B đều ngẫu nhiên)	Mô hình III (A cố định, B ngẫu nhiên)
Nhân tố A	$\frac{MS \text{ nhân tố B}}{MS \text{ ngẫu nhiên}}$	$\frac{MS \text{ nhân tố A}}{MS A \times B}$	$\frac{MS \text{ nhân tố A}}{MS A \times B}$
Nhân tố B	$\frac{MS \text{ nhân tố B}}{MS \text{ ngẫu nhiên}}$	$\frac{MS \text{ nhân tố B}}{MS A \times B}$	$\frac{MS \text{ nhân tố B}}{MS \text{ ngẫu nhiên}}$
A x B	$\frac{MS A \times B}{MS \text{ ngẫu nhiên}}$	$\frac{MS A \times B}{MS \text{ ngẫu nhiên}}$	$\frac{MS A \times B}{MS \text{ ngẫu nhiên}}$

5.5. Đa so sánh trong ANOVA hai nhân tố

Như đã đề cập trong chương 4, khi bác bỏ giả thuyết H_0 , chúng ta cần chỉ ra sự sai khác tồn tại giữa các nghiệm thức nào. Điều đó có nghĩa chúng ta cần tiến hành phân tích đa so sánh. Như đã đề cập trong chương 4, nếu các so sánh đã có kế hoạch từ trước và số lượng các so sánh không nhiều thì mỗi so sánh có thể dùng cùng một mức α mà đã được dùng trong phân tích phương sai. Nếu so sánh không có kế hoạch từ trước thì kiểm tra *Post hoc* như Scheffé hay Tukey hay LSD được dùng để điều khiển sai sót loại I.

Đa so sánh trong ANOVA hai nhân tố có cùng lý do, phương pháp như các đa so sánh trong ANOVA một nhân tố. Sự khác nhau duy nhất là: các giá trị $n_1, n_2, \dots$ trong các công thức của các kiểm tra nói trên chính là số lần lặp lại trong mỗi nhóm cần so sánh. Chúng ta có thể xem xét ví dụ trình bày ở bảng 5.6.

⁴ Thuật ngữ “Mô hình hỗn hợp - Mixed model” được giới thiệu bởi A.M. Mood vào năm 1950 (David, 1995).

Bảng 5.6. Ma trận các số liệu trong thí nghiệm nhân tố với 2×3 sự kết hợp

		Nhân tố B			Tổng
		B1	B2	B3	
Nhân tố A	A1	1	11	10	
		6	2	7	
		1	6	7	
		2	4	13	$\Sigma A_1 = 70$
		$\Sigma A_1 B_1 = 10$	$\Sigma A_1 B_2 = 23$	$\Sigma A_1 B_3 = 37$	$bn = 12$
		$n = 4$	$n = 4$	$n = 4$	
	A2	4	9	13	
		11	9	10	
		12	9	7	
		10	8	15	$\Sigma A_2 = 117$
		$\Sigma A_2 B_1 = 37$	$\Sigma A_2 B_2 = 35$	$\Sigma A_2 B_3 = 45$	$bn = 12$
		$n = 4$	$n = 4$	$n = 4$	
		$\Sigma B_1 = 47$	$\Sigma B_2 = 58$	$\Sigma B_3 = 82$	$\Sigma X_T = 187$
		$an = 8$	$an = 8$	$an = 8$	$N_T = 24$

Áp dụng quy trình phân tích phương sai như đã trình bày ở mục 5.3, kết quả phân tích phương sai được thể hiện ở bảng 5.7.

Bảng 5.7. Tóm tắt kết quả phân tích ANOVA từ số liệu ở bảng 5.6

Nguồn biến động	SS	df	MS	F
Nhân tố A	92,04	1	92,04	10,18**
Nhân tố B	80,08	2	40,04	4,43*
A x B	25,09	2	12,545	1,39
Ngẫu nhiên (trong các nhóm)	162,75	18	9,04	
Tổng	359,96	23		

Kết quả ở bảng 5.7 cho thấy chúng ta bác bỏ giả thuyết H_0 đối với ảnh hưởng của nhân tố B. Tuy nhiên, nhân tố B có 3 mức, chúng ta cần chỉ ra sự sai khác tồn tại giữa mức nào và mức nào.

5.5.1. Các so sánh có kế hoạch

Giả sử rằng hai so sánh trực giao sau đây là có kế hoạch từ trước khi thu thập số liệu từ trước B_1 với B_2 , trung bình của B_1 và B_2 với B_3 . Hệ số cân đối của các giá trị trung bình trong hai so sánh như sau:

So sánh	B ₁	B ₂	B ₃
B ₁ với B ₂	1	- 1	0
Trung bình của B ₁ và B ₂ với B ₃	1	1	- 2

Để kiểm tra mỗi so sánh, chúng ta tính giá trị $F_{\text{tính toán}}$ bằng công thức sau:

$$F_{\text{tính toán}} = \frac{C^2}{MS_W \left(\frac{w_1^2}{n_1} + \frac{w_2^2}{n_2} + \dots + \frac{w_t^2}{n_t} \right)} \quad [5.17]$$

Công thức tính cho ví dụ trên:

$$F_{\text{tính toán}} = \frac{C^2}{MS_W \left(\frac{w_1^2}{n_1} + \frac{w_2^2}{n_2} + \frac{w_3^2}{n_3} \right)} \quad [5.18]$$

Trong mỗi so sánh, C chính là hệ số tương phản. Như đề cập trong chương 4, $C = w_1M_1 + w_2M_2 + w_3M_3$; n_1, n_2, n_3 là số lần lặp lại trong các mức của nhân tố B, $n_1 = n_2 = n_3 = 8$; MS_W chính là 9,04. Ba giá trị trung bình ứng với ba mức của nhân tố B là $B_1 = 5,875$; $B_2 = 7,25$ và $B_3 = 10,25$.

Sau đó chúng ta đánh giá F bằng cách so sánh giá trị $F_{\text{tính toán}}$ với giá trị $F_{\text{quyết định}}$ với $df = (1, df_W)$. Bởi vì nhân tố B có ý nghĩa ở mức 0,05 cho nên ở đây ta dùng $\alpha = 0,05$ và $df_W = 18$.

- So sánh của B₁ với B₂

$$C = w_1M_1 + w_2M_2 + w_3M_3 = (1)(5,875) + (+1)(7,25) + (0)(10,25) = -1,375$$

$$F = \frac{(-1,375)^2}{9,04 \left(\frac{1^2}{8} + \frac{-1^2}{8} + \frac{0^2}{8} \right)} = 0,84$$

Bởi vì $F_{\text{tính toán}} < 1$, cho nên sự sai khác về tính trạng nghiên cứu giữa mức B₁ và B₂ không có ý nghĩa ở mức 0,05.

- So sánh B₁ và B₂ với B₃

$$C = w_1M_1 + w_2M_2 + w_3M_3 = (1)(5,875) + (-1)(7,25) + (-2)(10,25) = -7,375$$

$$F = \frac{(-7,375)^2}{9,04 \left(\frac{1^2}{8} + \frac{1^2}{8} + \frac{-2^2}{8} \right)} = 8,02$$

Với $df = (1, 18)$ giá trị $F_{\text{quyết định}}$ ở mức $\alpha = 0,05$ và $0,01$ là 4,41 và 8,28; như vậy giá trị F nằm giữa hai giá trị quyết định, cho nên chúng ta kết luận rằng trung bình của biến phụ thuộc tại mức trung bình của B₁ và B₂ khác với tại mức B₃.

5.5.2. Kiểm tra Scheffé

Giả sử không có so sánh nào có kế hoạch từ trước, nhưng chúng ta muốn so sánh giá trị trung bình ở mức (B_1 với B_3) và (B_1 và B_2 với B_3). Do các so sánh không có kế hoạch từ trước nên tiến hành kiểm tra *Post hoc* là thích hợp.

Trong kiểm tra Scheffé, các giá trị $F_{\text{tính toán}}$ được tính cho mỗi so sánh bằng cách sử dụng công thức giống như các so sánh có kế hoạch. Giá trị $F_{\text{quyết định}}$ được tính bằng cách nhân giá trị quyết định F ở mức ý nghĩa α trong phân tích ANOVA (trong trường hợp này là 0,05) với $k - 1$, trong đó k là số lượng trung bình cần được so sánh. Trong ANOVA, giá trị F với $\alpha = 0,05$ và $df(2, 18)$ là 3,55. Bởi vì có 3 mức của nhân tố B , giá trị quyết định khi $\alpha = 0,05$ trong kiểm tra Scheffé là $(3 - 1)(3,55) = 7,1$.

Trong kiểm tra B_1 với B_3 thì $w_1 = 1$, $w_2 = 0$ và $w_3 = -1$, các giá trị trung bình

$B_1 = 5,875$, $B_2 = 7,25$ và $B_3 = 10,25$, ta có:

$$C = w_1M_1 + w_2M_2 + w_3M_3 = (1)(5,875) + (0)(7,25) + (-1)(10,25) = -4,375$$

$$F = \frac{(-4,375)^2}{9,04\left(\frac{1^2}{8} + \frac{0^2}{8} + \frac{-1^2}{8}\right)} = 8,469$$

Như vậy, $F_{\text{tính toán}} > F_{\text{quyết định}}$ hay $8,45 > 7,1$ cho nên chúng ta kết luận rằng giá trị của biến phụ tại mức B_3 cao hơn mức B_1 .

Trường hợp so sánh giữa trung bình của (B_1 và B_2) với B_3 , giá trị $F_{\text{tính toán}}$ như đã tính ở trên là 8,02 và giá trị $F_{\text{quyết định}}$ là 7,1 cho nên chúng ta cũng quyết định rằng giá trị của biến phụ thuộc cao hơn ở mức B_3 so với sự kết hợp giữa B_1 và B_2 .

5.5.3. Kiểm tra Tukey HSD

Nếu nhóm nghiên cứu muốn so sánh mỗi cặp trong các mức độ của nhân tố B , kiểm tra Tukey là một lựa chọn thích hợp. Như đã đề cập trong chương 4, kiểm tra HSD có thể được thực hiện theo công thức [5.19].

$$HSD = q_{\text{quyết định}} \sqrt{\frac{MS_W}{n}} \quad [5.19]$$

HSD là sai khác trung thực có ý nghĩa giữa các cặp trung bình, n là số lần lặp lại trong mỗi nghiệm thức, $q_{\text{quyết định}}$ được tra từ bảng Tukey với mức ý nghĩa α , k (số trung bình cần kiểm tra) và df_W là độ tự do của ngẫu nhiên (lấy từ phân tích ANOVA). Trong ví dụ này $q_{\text{quyết định}}$ ở $\alpha = 0,05$, ba trung bình cần so sánh và $df_W = 18$ là 9,04.

$$HSD = 3,61 \sqrt{\frac{9,04}{8}} = 3,84$$

Sự khác nhau giữa các trung bình bằng 3,84 hay lớn hơn đều có ý nghĩa ở mức $\alpha = 0,05$. Sự khác nhau giữa B_2 và B_1 là 1,275; giữa B_3 và B_1 là 4,375; giữa B_3 và B_2 là 3,00. Như vậy, chỉ có sự khác nhau giữa B_1 với B_3 vượt quá giá trị HSD và do vậy có ý nghĩa ở mức α .

5.5.4. Kiểm tra Fisher LSD

Giống như kiểm tra Tukey HSD, kiểm tra Fisher LSD (sai khác ít nhất có ý nghĩa) được dùng để so sánh các cặp trung bình với nhau. Một thuận lợi của LSD là nó không yêu cầu các dung lượng mẫu bằng nhau. Kiểm tra Fisher LSD dựa vào:

$$t_0 = \frac{M_i - M_j}{\sqrt{MS_w \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad [5.20]$$

Giá trị t_0 được so sánh với giá trị t trong bảng t với độ tự do $df = df_w = N - k$. Nếu giá trị t_0 lớn hơn giá trị t bảng ở mức 0,05 thì giả thuyết H_0 bị bác bỏ. Ta có thể sử dụng kiểm tra LSD để kiểm tra trong trường hợp B_1 và B_3 .

$$t_0 = \frac{3,857 - 10,25}{\sqrt{9,04 \left(\frac{1}{8} + \frac{1}{8} \right)}} = -2,91$$

với $df = N - k = 24 - 3 = 21$ và $\alpha = 0,05$ giá trị tuyệt đối của t nằm trong khoảng 2,831 với $P = 0,01$ và 3,135 với $P = 0,005$, như vậy sai khác giữa B_1 và B_3 có ý nghĩa ở mức 0,01.

5.6. Phân tích phương sai hai nhân tố với số lần lặp lại không giống nhau

Thiết kế thí nghiệm có số lần lặp lại giống nhau là rất lý tưởng nhưng nhiều khi điều này lại không đạt được trong thực tế. Một số kiểu bố trí thí nghiệm hai nhân tố với các loại lặp lại khác nhau được thể hiện ở biểu đồ 5.4.

5.6.1. Lặp lại theo tỷ lệ

Bố trí thí nghiệm hai nhân tố lặp lại được gọi là lặp lại theo tỷ lệ nếu số lượng các số liệu trong ô trong hàng i và trong cột j là:

$$n_{ij} = \frac{(\text{số số liệu trong hàng } i) \times (\text{số số liệu trong cột } j)}{N} \quad [5.21]$$

Trong đó: N là tổng số các số liệu trong tất cả các ô. Ví dụ: hình (c) trong biểu đồ 5.4 có hai số liệu trong hàng 3 và cột 1 và $(16)(9)/72 = 2$. Cách phân tích phương sai được thể hiện ở bảng 5.7.

		Nhân tố A				Tổng số số liệu/hàng
		A1	A2	A3	A4	
Nhân tố B	B1	xxx	xxx	xxx	xxx	12
	B2	xxx	xxx	xxx	xxx	12
	B3	xxx	xxx	xxx	xxx	12
Tổng số số liệu/cột		9	9	9	9	N = 36

(a) Số lần lặp lại bằng nhau ở mỗi sự kết hợp giữa các nghiệm thức

		Nhân tố A				Tổng số số liệu/hàng
		A1	A2	A3	A4	
Nhân tố B	B1	xxx	xxx	xxx	xxx	12
	B2	xxxx	xxxx	xxxx	xxxx	16
	B3	xx	xx	xx	xx	8
Tổng số số liệu/cột		9	9	9	9	N = 36

(b) Số lần lặp lại bằng nhau trong các hàng; lặp lại theo tỷ lệ trong các cột

		Nhân tố A				Tổng số số liệu/hàng
		A1	A2	A3	A4	
Nhân tố B	B1	xxx	xxx	xxx	xxx	24
			xxx	xxx	xxx	
				xxx		
	B2	xxxx	xxxx	xxxx	xxxx	32
			xxxx	xxxx	xxxx	
				xxxx	xxxx	
	B3	xx	xx	xx	xx	16
			xx	xx	xx	
				xx		
Tổng số số liệu/cột		9	18	27	18	N = 72

(c) Lặp lại theo tỷ lệ trong các hàng và trong các cột

		Nhân tố A				Tổng số số liệu/hàng
		A1	A2	A3	A4	
Nhân tố B	B1	xxx	xx	xx	xxxx	11
	B2	xxxx	xx	xx	xxx	11
	B3	xxxx	xxx	xxxx	xxx	14
Tổng số số liệu/cột		11	7	8	10	N = 36

(d) Lặp lại không theo tỷ lệ

		Nhân tố A				Tổng số số liệu/hàng
		A1	A2	A3	A4	
Nhân tố B	B1	x	x	x	x	4
	B2	x	x	x	x	4
	B3	x	x	x	x	4
Tổng số số liệu/cột		3	3	3	3	N = 12

(e) Không lặp lại

Biểu đồ 5.4. Các loại bố trí thí nghiệm hai nhân tố có số lần lặp lại khác nhau

Bảng 5.8. ANOVA hai nhân tố với số lần lặp lại theo tỷ lệ

Nguồn biến động	SS	df	MS
Tổng số	$\sum X_T^2 - \frac{(\sum X_T)^2}{N_T}$	N - 1	
Giữa các nhóm	$\frac{(\sum A_1 B_1)^2}{n_{11}} + \frac{(\sum A_1 B_2)^2}{n_{12}} + \dots + \frac{(\sum A_b B_b)^2}{n_{ab}} - \frac{(\sum X_T)^2}{N_T}$	ab - 1	MS _{giữa các nhóm/ df_{giữa các nhóm}}
Nhân tố A	$\frac{(\sum A_1)^2}{\sum_{j=1}^b n_{1j}} + \frac{(\sum A_2)^2}{\sum_{j=1}^b n_{2j}} + \dots + \frac{(\sum A_a)^2}{\sum_{j=1}^b n_{aj}} - \frac{(\sum X_T)^2}{N_T}$	a - 1	$\frac{SS_A}{df_A}$
Nhân tố B	$\frac{(\sum B_1)^2}{\sum_{j=1}^a n_{1j}} + \frac{(\sum B_2)^2}{\sum_{j=1}^a n_{2j}} + \dots + \frac{(\sum B_b)^2}{\sum_{j=1}^a n_{bj}} - \frac{(\sum X_T)^2}{N_T}$	b - 1	$\frac{SS_B}{df_B}$
A x B	SS _{giữa các nhóm} - SS _{nhân tố A} - SS _{nhân tố B}	(a - 1)(b - 1)	$\frac{SS_{A \times B}}{df_{A \times B}}$
Ngẫu nhiên	SS _T - SS _{giữa các nhóm}	df _T - df _{giữa các nhóm}	$\frac{SS_W}{df_W}$

5.6.2. Lặp lại không theo tỷ lệ

Nếu một số ô/sự kết hợp giữa các nghiệm thức có số lượng các số liệu/số lần lặp lại vượt quá sự cần thiết cho bố trí thí nghiệm cân bằng hoặc theo tỷ lệ thì khi đó các số liệu có thể được bỏ bớt một cách ngẫu nhiên để trở thành một thí nghiệm cân bằng hoặc theo tỷ lệ. Và khi đó ANOVA được tiến hành một cách bình thường như đã đề cập. Nếu một ô thiếu một số liệu, để đảm bảo cho cân bằng theo tỷ lệ thì có thể ước tính giá trị đó theo phương pháp của Shearer (1973).

$$\hat{X}_{ijl} = \frac{aA_i + bB_j - \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^{n_{ij}} X_{ijl}}{N+1-a-b}$$

Trong đó

- $\hat{X}_{ijl}$ là giá trị ước tính cho lần lặp lại l trong mức i của nhân tố A và mức j của nhân tố B
- A_i là tổng các số liệu trong mức i của nhân tố A
- B_j là tổng số các số liệu trong mức j của nhân tố B
- $\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^{n_{ij}} X_{ijl}$ là tổng số các số liệu đã biết
- N là tổng số các số liệu kể cả số liệu bị mất hoặc không có

Ví dụ: Nếu một số liệu X_{124} trong ví dụ ở bảng 5.2 không có chúng ta có thể ước tính giá trị này: $a = 2$, $b = 2$, $N = 20$, $A_1 = A_1$ = tổng số các số liệu đã biết từ các cá thể không có hormon, $B_2 = B_2$ = tổng số các số liệu đã biết từ các con trống và $\sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^{n_{ij}} X_{ijl}$ = tổng số của 19 số liệu đã biết từ cả có và không có hormon và cả hai giới tính. Sau khi ước tính giá trị bị mất ta điền giá trị đó vào vị trí bị mất và tiến hành ANOVA. Tuy nhiên khi xem xét giá trị df_T và df_W thì giá trị ước tính đó không được xem xét có nghĩa là $df_T = 18$ và $df_W = 15$.

Nếu một ô thiếu nhiều hơn một số liệu (nhưng không vượt quá 10% tổng số số liệu hay nhiều số liệu thiếu hơn số mức độ của bất cứ nhân tố nào), khi đó phương trình trên có thể được sử dụng lặp đi lặp lại để tính các giá trị thiếu. Số lượng các ước tính đó không nên được đưa vào df_T hoặc df_W .

5.6.3. Phân tích phương sai hai nhân tố không có lần lặp lại

Bảng 5.9. Phân tích ANOVA hai nhân tố không có lặp lại

Nguồn biến động	SS	df	MS
Tổng số	$\sum X_T^2 - \frac{(\sum X_T)^2}{N_T}$	N - 1	
Nhân tố A	$\frac{(\sum A_1)^2}{b} + \frac{(\sum A_2)^2}{b} + \dots + \frac{(\sum A_a)^2}{b} - \frac{(\sum X_T)^2}{N_T}$	a - 1	$\frac{SS_A}{df_A}$
Nhân tố B	$\frac{(\sum B_1)^2}{a} + \frac{(\sum B_2)^2}{a} + \dots + \frac{(\sum B_b)^2}{a} - \frac{(\sum X_T)^2}{N_T}$	b - 1	$\frac{SS_B}{df_B}$
Ngẫu nhiên	$SS_T - SS_{\text{nhân tố A}} - SS_{\text{nhân tố B}}$	(a - 1)(b - 1)	$\frac{SS_{\text{còn lại}}}{df_{\text{còn lại}}}$

Đôi khi ta gặp trường hợp bố trí thí nghiệm hai nhân tố mà chỉ có một số liệu cho mỗi sự kết hợp của nhân tố ($n = 1$ cho tất cả các ô). Trong trường hợp như vậy, mỗi số liệu có thể được ký hiệu là X_{ij} , trong đó i là một mức độ nào đó của nhân tố A và j là một mức độ nào đó của nhân tố B.

Phương pháp phân tích ANOVA trong trường hợp này được thể hiện ở bảng 5.9. Khi $n = 1$, SS giữa các nhóm/ô/sự kết hợp giữa các mức của nhân tố là trùng với SS_T và df giữa các nhóm là trùng với df_T , do vậy SS_W và df_W sẽ bằng 0 có nghĩa là với một số liệu ở mỗi kết hợp thì không có sự biến động trong các ô.

Ngoài ảnh hưởng của nhân tố A và B, trong SS_T còn có một phần nữa: $SS_{\text{còn lại}} = SS_T - SS_A - SS_B$ với $df_{\text{còn lại}} = df_T - df_A - df_B$. Phần này chính là lượng tương tác giữa hai nhân tố A và B.

Các mô hình phân tích ANOVA trong trường hợp bố trí thí nghiệm nhân tố không có lặp lại được thể hiện ở bảng 5.10.

Bảng 5.10. Giá trị F tính toán trong ANOVA hai nhân tố không lặp lại

Ảnh hưởng	Mô hình I (cả A và B đều cố định)	Mô hình II (cả A và B đều ngẫu nhiên)	Mô hình III (A cố định, B ngẫu nhiên)
<i>Giả thiết rằng có ảnh hưởng có ý nghĩa của tương tác</i>			
Nhân tố A	Kiểm tra cẩn thận*	$\frac{MS \text{ nhân tố A}}{MS \text{ còn lại}}$	$\frac{MS \text{ nhân tố A}}{MS \text{ còn lại}}$
Nhân tố B	Kiểm tra cẩn thận*	$\frac{MS \text{ nhân tố B}}{MS \text{ còn lại}}$	Kiểm tra cẩn thận*
A x B	Không thể kiểm tra	Không thể kiểm tra	Không thể kiểm tra
<i>Không có ảnh hưởng có ý nghĩa của tương tác</i>			
Nhân tố A	$\frac{MS \text{ nhân tố A}}{MS \text{ còn lại}}$	$\frac{MS \text{ nhân tố A}}{MS \text{ còn lại}}$	$\frac{MS \text{ nhân tố A}}{MS \text{ còn lại}}$
Nhân tố B	$\frac{MS \text{ nhân tố B}}{MS \text{ còn lại}}$	$\frac{MS \text{ nhân tố B}}{MS \text{ còn lại}}$	$\frac{MS \text{ nhân tố B}}{MS \text{ còn lại}}$
A x B	Không thể kiểm tra	Không thể kiểm tra	Không thể kiểm tra

Bởi vì chúng ta không có các đo lường sự biến động của sự tương tác giữa hai nhân tố và sai sót ngẫu nhiên cho nên kiểm tra ảnh hưởng của mỗi nhân tố trong mô hình ANOVA I hay kiểm tra ảnh hưởng nhân tố ngẫu nhiên trong mô hình ANOVA III là không nên, bởi vì có thể có tương tác giữa hai nhân tố.

5.7. Bài tập

1. Hãy giải thích tại sao tiến hành một thí nghiệm nhiều nhân tố có hiệu quả thiết kế hơn so với tiến hành nhiều thí nghiệm riêng lẻ ứng với mỗi thí nghiệm một nhân tố?
2. Một thí nghiệm nghiên cứu ảnh hưởng của nhân tố B (gồm có 3 mức B_1 , B_2 và B_3) và nhân tố A (gồm có 2 mức A_1 và A_2) đến giá trị của tính trạng Y. Thí nghiệm được thiết kế theo kiểu CRD với tổng số cá thể nghiên cứu (số đơn vị thí nghiệm) là 24. Kết quả giá trị của tính trạng Y được trình bày ở bảng sau.

		Nhân tố B		
		B1	B2	B3
A1	1	11	10	
	6	2	7	
	1	6	7	
	2	4	13	
A2	4	9	13	
	11	9	10	
	12	9	7	
	10	8	15	

Hãy:

- a. Viết mô hình thống kê phân tích ảnh hưởng của nhân tố nghiên cứu A và B đến giá trị của tính trạng Y.
 - b. Hãy thiết kế mẫu nhập số liệu để có thể phân tích ảnh hưởng của nhân tố nghiên cứu A và B đến giá trị của tính trạng y bằng các phần mềm thống kê thông dụng như Minitab, Genstat, hay SPSS.
 - c. Hãy tiến hành phân tích phương sai để kiểm tra giả thuyết về ảnh hưởng của nhân tố nghiên cứu A.
 - d. Hãy tiến hành phân tích *Post hoc* về ảnh hưởng của các mức của nhân tố B đến tính trạng y bằng cách sử dụng cả ba loại kiểm tra sau: Scheffé, Tukey HSD và Fisher LSD.
 - e. Hãy thực hiện so sánh có kế hoạch về ảnh hưởng lên tính trạng Y của sự kết hợp của B_1 và B_2 với B_3 .
3. Một nhóm nghiên cứu ảnh hưởng của 2 mức protein thô (CP1 và CP2) và 3 mức carbohydrate phi tinh bột (NSP1; NSP2; NSP3) trong khẩu phần đến tăng trọng của lợn. Thí nghiệm được tiến hành trên 36 con lợn theo kiểu thiết kế CRD. Thời gian thí nghiệm kéo dài 95 ngày. Khối lượng ban đầu (kg) và khối lượng kết thúc (kg) của lợn được thể hiện ở bảng sau. Hãy phân tích phương sai kiểm tra ảnh hưởng của mức protein và mức NSP trong khẩu phần đến tăng trọng (g/ngày) của lợn.

Mức protein	Mức NSP	Khối lượng ban đầu	Khối lượng kết thúc
CP1	NSP1	30,3	104,2
CP1	NSP1	27,6	103,7
CP1	NSP1	24,1	103,4
CP1	NSP1	25,3	101,2
CP1	NSP1	26,1	107,8
CP1	NSP1	25,6	102,4
CP1	NSP2	25,6	110,5
CP1	NSP2	26,6	106,7
CP1	NSP2	24,5	102,4
CP1	NSP2	23,2	101,9
CP1	NSP2	25,9	112,3
CP1	NSP2	23,5	107,2
CP1	NSP3	28,4	105,2
CP1	NSP3	25,2	104,3
CP1	NSP3	22,3	100,8
CP1	NSP3	20,7	103,5
CP1	NSP3	26,9	107,3
CP1	NSP3	25,1	104,9
CP2	NSP1	27,4	101,9
CP2	NSP1	24,9	104,8
CP2	NSP1	23,9	100,2
CP2	NSP1	22,5	104,7
CP2	NSP1	28,3	108,8
CP2	NSP1	20,9	105,4
CP2	NSP2	26	108,9
CP2	NSP2	25,9	106,2
CP2	NSP2	22,7	105,4
CP2	NSP2	21,7	102,8
CP2	NSP2	27,8	109,8
CP2	NSP2	22,6	100,5
CP2	NSP3	28	109,9
CP2	NSP3	27,1	107,5
CP2	NSP3	30,3	105,7
CP2	NSP3	23,2	97
CP2	NSP3	26,5	97
CP2	NSP3	22,2	102,8

Chương 6

THIẾT KẾ THÍ NGHIỆM MỘT NHÂN TỐ

Chương này sẽ đề cập đến 3 kiểu thiết kế thí nghiệm cơ bản thường gặp:

1. *Thiết kế thí nghiệm ngẫu nhiên hoàn toàn (CRD)*
2. *Thiết kế thí nghiệm khối ngẫu nhiên đầy đủ (RCB)*
3. *Thiết kế thí nghiệm ô vuông la tinh (LSD)*

Một số kiểu thiết kế thí nghiệm khác ví dụ: thiết kế thí nghiệm ô lớn ô nhỏ; một số dạng biến thể của thiết kế thí nghiệm kiểu RCB và LSD như khối ngẫu nhiên không đầy đủ sẽ không được đề cập trong nội dung của giáo trình này. Trong mỗi kiểu thiết kế thí nghiệm, các nội dung chính được trình bày gồm:

- a. *Đặc điểm kiểu thiết kế thí nghiệm*
- b. *Cách thiết kế*
- c. *Đơn vị thí nghiệm*
- d. *Số lần lặp lại*
- e. *Mô hình phân tích thống kê*
- f. *Chuẩn bị số liệu, phân tích số liệu và diễn dịch kết quả*

6.1. Thiết kế thí nghiệm ngẫu nhiên hoàn toàn

6.1.1. Định nghĩa

Kiểu thiết kế thí nghiệm ngẫu nhiên hoàn toàn là kiểu thiết kế mà các nghiệm thức được bố trí một cách hoàn toàn ngẫu nhiên vào các đơn vị thí nghiệm hay các đơn vị thí nghiệm được bố trí một cách hoàn toàn ngẫu nhiên vào các nghiệm thức.

6.1.2. Mô tả thiết kế thí nghiệm ngẫu nhiên hoàn toàn

Kiểu thiết kế thí nghiệm ngẫu nhiên hoàn toàn được ký hiệu là CRD - Completely Randomized Design. Khái niệm hoàn toàn ngẫu nhiên ở đây có nghĩa là các đơn vị thí nghiệm đều có cơ hội/xác suất ngang nhau được bố trí vào các nghiệm thức và ngược lại. Kiểu thiết kế thí nghiệm CRD có một số thuận lợi như sau:

- *Tính linh hoạt cao:* Tính linh hoạt được thể hiện trên khía cạnh số nghiệm thức và số lần lặp lại trong mỗi nghiệm thức. Số lần lặp lại có thể khác nhau ở các nghiệm thức khác nhau, mặc dù sự khác nhau này không được khuyến

khích nếu không có lý do chính đáng. Trường hợp lý tưởng nhất vẫn là các nghiệm thức có số lần lặp lại giống nhau. Điều kiện này góp phần đảm bảo tính đồng nhất về phương sai khi phân tích ANOVA như đã đề cập trong chương 4.

- *Phân tích thống kê dễ dàng:* Kiểu thiết kế thí nghiệm CRD có mô hình phân tích thống kê đơn giản nhất trong tất cả các kiểu thiết kế thí nghiệm, thậm chí cả khi số lần lặp lại ở các nghiệm thức khác nhau.
- *Phương pháp phân tích* vẫn đơn giản khi kết quả của một vài đơn vị thí nghiệm thậm chí của cả nghiệm thức bị mất. Hơn thế nữa ảnh hưởng của việc mất thông tin do mất số liệu đến kết quả diễn dịch là ít hơn so với các kiểu thiết kế thí nghiệm khác.
- *Độ tự do của sai số cao nhất:* Với bất kỳ số nghiệm thức và số lần lặp lại nào thì kiểu thiết kế thí nghiệm CRD có độ tự do của sai số (df_w) cao nhất. Lý do là kiểu thiết kế thí nghiệm này không bị khống chế bởi bất kỳ yếu tố nào trong quá trình ngẫu nhiên hóa. Mỗi sự khống chế trong thiết kế đều gây tiêu tốn độ tự do cho nên làm giảm df_w . Độ nhạy của thí nghiệm được tăng lên khi df_w tăng và ngược lại. Độ nhạy của thí nghiệm có nghĩa là khả năng bác bỏ giả thuyết H_0 khi giả thuyết H_0 sai và khả năng chấp nhận giả thuyết H_0 khi giả thuyết H_0 đúng.

Tuy vậy, độ chính xác của kiểu thiết kế thí nghiệm CRD thường được cân nhắc và độ chính xác cũng là lý do làm khống chế việc sử dụng kiểu thiết kế thí nghiệm này. Quá trình ngẫu nhiên hóa các nghiệm thức vào các đơn vị thí nghiệm và ngược lại trong kiểu thiết kế thí nghiệm này không bị khống chế bởi bất cứ yếu tố nào. Điều này có nghĩa tất cả biến động giữa các đơn vị thí nghiệm đều được hội tụ ở sai số của thí nghiệm, mà ước tính phương sai là MS_w . Trong khi đó, sai số thí nghiệm có thể được khống chế bởi các kiểu thiết kế thí nghiệm khác như RCB và LSD.

Kiểu thiết kế thí nghiệm CRD được sử dụng khi các đơn vị thí nghiệm tương đồng về các yếu tố mà có thể ảnh hưởng đến biến phụ thuộc. Ví dụ, để nghiên cứu ảnh hưởng của các mức protein trong khẩu phần đến khả năng tăng trọng, hiệu quả chuyển hóa thức ăn và phẩm chất thịt xẻ của lợn lai F1 (Móng Cái $\times$ Đại Bạch) thì các đơn vị thí nghiệm phải tương đồng về tuổi, khối lượng ban đầu, giới tính. Khi xuất phát điểm của các đơn vị thí nghiệm là không tương đồng về các tính trạng có thể ảnh hưởng đến tính trạng nghiên cứu thì chúng ta không nên dùng kiểu thiết kế thí nghiệm CRD.

Kiểu thiết kế thí nghiệm CRD thường được dùng trong các thí nghiệm được tiến hành trong phòng thí nghiệm và thông thường được các lĩnh vực lý hóa học sử dụng nhiều. Trong khi đó các thí nghiệm trên đồng ruộng thường ít sử dụng kiểu thiết kế thí nghiệm này. Trong lĩnh vực nông nghiệp, lĩnh vực khoa học vật nuôi thường sử dụng kiểu thiết kế thí nghiệm này hơn là lĩnh vực khoa học cây trồng. Khi hệ thống sản xuất càng đa dạng, càng hỗn hợp thì kiểu thiết kế thí nghiệm này càng ít được sử dụng.

6.1.3. Ngẫu nhiên hóa

6.1.3.1. Sự cần thiết của tính ngẫu nhiên hóa trong thiết kế thí nghiệm

Như đã đề cập trong chương trước, tính lặp lại và tính ngẫu nhiên là hai yêu cầu không thể thiếu được trong bất kỳ thí nghiệm nào. Một số thí nghiệm còn yêu cầu khống chế sự sai khác ban đầu của đơn vị thí nghiệm. Lặp lại cung cấp số liệu cho việc ước tính phương sai của sai số thí nghiệm. Khống chế sự sai khác ban đầu, ví dụ sử dụng kỹ thuật thiết kế theo khối, là một cách thức để khống chế sai số thí nghiệm. Tuy nhiên cả yêu cầu lặp lại và khống chế sự sai khác ban đầu không thể bảo đảm ước tính phương sai của sai số thí nghiệm một cách hợp lệ. Trong khi đó ngẫu nhiên hóa là điều kiện để ước tính phương sai của sai số thí nghiệm một cách hợp lệ, phục vụ cho suy diễn thống kê.

Phân tích ANOVA yêu cầu các quan sát độc lập và là một mẫu ngẫu nhiên xuất phát từ một quần thể phân bố chuẩn. Để các quan sát ngẫu nhiên và độc lập, đòi hỏi các đơn vị thí nghiệm được bố trí một cách ngẫu nhiên vào các nghiệm thức. Để có thể so sánh ảnh hưởng của các nghiệm thức cần thiết phải xóa bỏ sự thiên lệch trong việc bố trí một cách có hệ thống các đơn vị thí nghiệm vào các nghiệm thức. Ví dụ, chúng ta muốn so sánh hai phương pháp phân tích (A và B) hàm lượng H_2S trong không khí phát xạ từ phân lợn. Chúng ta sử dụng phương pháp A để xác định trên một số mẫu không khí và phương pháp B trên một số mẫu không khí khác. Phương pháp A do kỹ thuật viên thứ nhất thực hiện, phương pháp B do kỹ thuật viên thứ 2 thực hiện. Nếu chúng ta thực hiện như vừa nêu thì sẽ tạo nên biến động mang tính hệ thống do kỹ thuật viên gây nên và nguyên tắc ngẫu nhiên không được đảm bảo. Chúng ta sẽ không thể có được so sánh chính xác giữa hai phương pháp, vì kết quả phân tích còn có thể bị ảnh hưởng bởi kỹ thuật viên.

6.1.3.2. Phương pháp ngẫu nhiên

Ngẫu nhiên hóa trong thiết kế thí nghiệm được hiểu một cách đầy đủ là quá trình sắp xếp một tập hợp các đối tượng theo một trật tự ngẫu nhiên. Các đối tượng đó có thể là các nghiệm thức, các khối, các cột hoặc các hàng... trong các kiểu thiết kế. Phương pháp truyền thống để bố trí một cách ngẫu nhiên các đơn vị thí nghiệm vào các nghiệm thức là sử dụng ném đồng xu hoặc tung con xúc xắc. Phương pháp này dần dần được thay thế bởi bảng số ngẫu nhiên được tạo sẵn ở các tài liệu về thống kê và phương pháp thí nghiệm. Ngoài ra các số ngẫu nhiên còn có thể được tạo ra bởi hàm ngẫu nhiên [Rand() hoặc Randbetween()] trong các phần mềm tin học như Genstat, Minitab, SPSS hay Excel. Phương pháp ngẫu nhiên hóa bằng cách sử dụng các hàm ngẫu nhiên được khuyến khích vì sự tiện ích của nó trong quá trình thiết kế thí nghiệm.

Chúng ta hãy xem xét ví dụ 4.3 trong chương 4.

“Ví dụ 4.3. Ảnh hưởng của các mức protein thô (CP) trong khẩu phần (13%, 15% và 17%) đến tăng trọng (kg/tháng) của lợn lai F1 (Móng Cái × Đại Bạch) được bộ môn

Dinh dưỡng và Thức ăn, khoa Chăn nuôi Thú y, Đại học Nông lâm Huế nghiên cứu, kết quả được trình bày ở bảng 4.1. Thí nghiệm được tiến hành trên 24 con lợn (N), số lần lặp lại ở mỗi nghiệm thức CP là 8 con lợn (n). Giả sử tăng trọng của lợn ở mỗi nghiệm thức CP đáp ứng yêu cầu của phân tích phương sai”.

Thí nghiệm được tiến hành trên 24 con lợn (N) tương đồng về các yếu tố khối lượng, tuổi tác, giới tính, giống... Lợn được nuôi cá thể trong các ô chuồng. Như đã đề cập ở trên với điều kiện tương đồng của các đơn vị thí nghiệm, chúng ta nên sử dụng kiểu thiết kế thí nghiệm CRD. Với 24 cá thể và 3 nghiệm thức, sẽ có 8 lần lặp lại cho mỗi nghiệm thức.

Phương pháp ngẫu nhiên hóa có thể được tiến hành theo các bước sau và được thực hiện trong một trang bảng tính (sheet) của Excel:

Bước 1 (B1): Lập danh sách (24) đơn vị thí nghiệm (mỗi con lợn hay mỗi ô chuồng) và đánh số các đơn vị thí nghiệm theo thứ tự (từ 1 - 24)

Bước 2 (B2): Gán cho mỗi đơn vị thí nghiệm một số ngẫu nhiên bằng cách sử dụng hàm RAND(). Hàm RAND() cho phép tạo ra các số ngẫu nhiên có giá trị trong khoảng [0 - 1]. Chúng ta có thể tạo ra các số ngẫu nhiên trong một khoảng giá trị nào đó bằng cách sử dụng hàm RANDBEETWEEN(giá trị 1; giá trị 2). Điều cần đặc biệt quan tâm là kết quả hàm RAND() luôn luôn thay đổi, do vậy chúng ta phải cố định các giá trị ngẫu nhiên ngay sau khi đã tạo ra. Quá trình cố định này có thể thực hiện thông qua trình ứng dụng Paste special - Paste value trong phần mềm Excel. Cần dán (paste) các giá trị ngẫu nhiên vừa tạo ra trên chính các giá trị vừa tạo ra. Mục đích là chuyển từ dạng công thức sang dạng giá trị.

Bước 3 (B3): Gán các nghiệm thức vào các đơn vị thí nghiệm thứ tự (từ 13%CP đến 17%CP) (Bảng 6.1)

Bước 4 (B4): Sắp xếp các cột “số ngẫu nhiên” và “mức protein” theo thứ tự tăng dần hoặc giảm dần của “số ngẫu nhiên”.

Kết quả cho thấy các con lợn nhận nghiệm thức 13%CP là: 2; 6; 9; 13; 14; 16; 18 và 21. Những con lợn sau nhận nghiệm thức 15%CP là: 1; 10; 15; 19; 20; 22; 23 và 24. Các con lợn còn lại nhận nghiệm thức 17%CP. Cần chú ý rằng mỗi lần thực hiện bố trí ngẫu nhiên các đơn vị thí nghiệm vào các nghiệm thức luôn cho kết quả khác nhau, do hàm ngẫu nhiên luôn luôn cho kết quả các số ngẫu nhiên khác nhau.

Một số tác giả ví dụ Kehl (2003) gọi cách ngẫu nhiên vừa trình bày ở trên là ngẫu nhiên đơn. Ngẫu nhiên đơn được đặc trưng bởi thực hiện lệnh Sort (sắp xếp) một lần. Cách gọi ngẫu nhiên đơn nhằm phân biệt với ngẫu nhiên kép. Trong phương pháp ngẫu nhiên kép chúng ta có thể thực hiện ngẫu nhiên hóa các con lợn vào các ô chuồng bằng cách Sort cột 1 và cột 2 theo cột 2. Sau khi thực hiện B3 chúng ta tiến hành Sort một lần nữa nhưng Sort theo cột 3.

Bảng 6.1. Quá trình ngẫu nhiên hóa đơn trong thiết kế thí nghiệm kiểu CRD

B1			B2			B3			B4		
Cột 1			Cột 2			Cột 3			Cột 1		
Con lợn			Số ngẫu nhiên			Thí nghiệm thức			Con lợn		
1			0,485241372			13%CP			1		
2			0,83296448			15%CP			2		
3			0,755481259			17%CP			3		
4			0,792772895			13%CP			4		
5			0,033928302			15%CP			5		
6			0,421183444			17%CP			6		
7			0,653933273			13%CP			7		
8			0,490003494			15%CP			8		
9			0,535638176			17%CP			9		
10			0,713850843			13%CP			10		
11			0,940457894			15%CP			11		
12			0,448307137			17%CP			12		
13			0,035714957			13%CP			13		
14			0,967915049			15%CP			14		
15			0,105024436			17%CP			15		
16			0,656308708			13%CP			16		
17			0,819146507			15%CP			17		
18			0,049174649			17%CP			18		
19			0,155558330			13%CP			19		
20			0,949684416			15%CP			20		
21			0,061979632			17%CP			21		
22			0,913293136			13%CP			22		
23			0,663214746			15%CP			23		
24			0,643343235			17%CP			24		

Trong trường hợp không có máy vi tính để tạo số ngẫu nhiên cũng như không có bảng số ngẫu nhiên chúng ta có thể thực hiện theo phương pháp tương đương sau:

Bước 1: Lập danh sách (24) đơn vị thí nghiệm và đánh số các đơn vị thí nghiệm theo thứ tự (từ 1 - 24). Giống như B1 trong cách bố trí ngẫu nhiên ở trên

Bước 2: Làm (24) cái thăm với các con số giống như các con số đánh cho các đơn vị thí nghiệm

Bước 3: Bóc n (8) thăm đầu tiên cho nghiệm thức thứ nhất, bóc n (8) thăm tiếp theo cho nghiệm thức thứ 2 và số còn lại là cho nghiệm thức thứ 3.

6.1.4. Mô hình thống kê và phân tích số liệu

6.1.4.1. Mô hình thống kê

Một điều đặc biệt lưu ý trước khi tiến hành thí nghiệm chính thức là phải xác định rõ mô hình thống kê phân tích số liệu để trả lời câu hỏi nghiên cứu. Khi chưa rõ mô hình thống kê thì chưa tiến hành thí nghiệm. Thậm chí phải chạy thử mô hình thống kê từ số liệu giả định trước khi tiến hành thí nghiệm. Vấn đề này càng đặc biệt chú ý hơn khi chúng ta có các kiểu thiết kế thí nghiệm phức tạp ví dụ, Box Benkin hay Boxgensen.

Mô hình thống kê cho thiết kế thí nghiệm kiểu CRD một nhân tố có dạng:

$$y_{ij} = \mu + T_i + e_{ij} \quad [6.1]$$

Trong đó:

- y_{ij} là giá trị quan sát (ở ví dụ 4.3 là tăng trọng, kg/tháng) thứ j từ nghiệm thức i
- μ là trung bình quần thể
- T_i là ảnh hưởng của nhân tố nghiên cứu (trong ví dụ 4.3 là mức protein)
- e_{ij} là sai số thí nghiệm
- i là mức của nhân tố nghiên cứu $i = 1, 2, 3, \dots, t$ (trong ví dụ 4.3, $i = 1, 2, 3$), t là tổng số mức của nhân tố nghiên cứu
- j là số lần lặp lại $= n$ (trong ví dụ 4.3 $n = 8$)

6.1.4.2. Chuẩn bị số liệu để phân tích

Sau khi đã xác định rõ các thông tin như đơn vị thí nghiệm, số lần lặp lại, mô hình thống kê, chúng ta tiến hành thí nghiệm và thu thập số liệu. Khi đã có số liệu chúng ta cần chuẩn bị số liệu cho việc xử lý thống kê mô tả và thống kê suy diễn. Chuẩn bị số liệu thường được thực hiện bởi các phần mềm tin học mà phổ biến nhất là Excel. Thông thường Excel dùng để quản lý số liệu, ít khi được sử dụng để phân tích số liệu, mặc dầu Excel vẫn có thể phân tích được thống kê mô tả và một số thống kê suy diễn. Trong nội dung của giáo trình này chúng tôi hạn chế Excel ở mức độ quản lý số liệu. Đa số các phần mềm thống kê thông dụng hiện nay như Genstat, NLG, SPSS, Minitab có cùng cách nhận biết số liệu từ Excel. Nguyên tắc cơ bản nhất của chuẩn bị số liệu trong Excel là “mỗi hàng là một đơn vị thí nghiệm và mỗi cột là một nghiệm thức”. Bảng 4.1 (chương 4) đã trình bày các số liệu thí nghiệm thu thập được. Để có thể xử lý được bởi các phần mềm thống kê vừa nêu, các số liệu được sắp xếp như sau:

Bảng 6.2. Chuẩn bị số liệu cho phân tích số liệu của thí nghiệm thiết kế theo kiểu CRD

Mức protein	Tăng trọng (kg/tháng)
13%CP	11
13%CP	3
13%CP	5
13%CP	10
13%CP	7
13%CP	8
13%CP	9
13%CP	2
15%CP	15
15%CP	14
15%CP	10
15%CP	12
15%CP	13
15%CP	6
15%CP	9
15%CP	10
17%CP	13
17%CP	14
17%CP	6
17%CP	10
17%CP	12
17%CP	10,5
17%CP	16
17%CP	13

6.1.4.3. Phương pháp phân tích số liệu

Căn cứ vào mô hình thống kê để quyết định phương pháp phân tích số liệu. Chúng ta cần phân tích hai loại thống kê cơ bản: thống kê mô tả và phân tích thống kê suy diễn. Thống kê mô tả thường đặc trưng bởi các tham số đặc trưng cho mức độ tập trung và tham số đặc trưng cho mức độ phân tán. Thông thường chúng ta sử dụng giá trị trung bình để đặc trưng cho mức độ tập trung. Khi các số liệu của biến phụ thuộc là phân bố chuẩn và đồng nhất phương sai thì ta sử dụng giá trị trung bình thực, còn khi không đảm bảo điều kiện đó thì sử dụng giá trị trung bình nhân. Độ lệch chuẩn thông thường là tham số đặc trưng cho mức độ phân tán, chúng ta có thể sử dụng hệ số biến dị cũng như các tham số khác. Phương pháp phân tích thống kê mô tả đã được đề cập trong chương 2. Kết quả phân tích thống kê mô tả về tăng trọng của lợn ở ví dụ 4.3 được trình bày ở bảng 6.3.

Bảng 6.3. Một số tham số thống kê mô tả về tăng trọng của lợn (kg/tháng) ở ví dụ 4.3

Tham số	13%CP	15%CP	17%CP
Trung bình	6,88	11,13	11,81
Sai số của số trung bình	1,16	1,04	1,07
Trung vị	7,50	11,00	12,50
Yếu số	NA	10,00	13,00
Độ lệch chuẩn	3,27	2,95	3,02
Phương sai	10,70	8,70	9,14
Khoảng biến động	9,00	9,00	10,00

Trong kiểu thiết kế CRD một nhân tố, để trả lời câu hỏi nghiên cứu có hay không có ảnh hưởng của nhân tố nghiên cứu đến các chỉ tiêu nghiên cứu chúng ta cần phân tích ANOVA một nhân tố như đã đề cập ở chương 4. Kết quả phân tích phương sai được trình bày dưới dạng bảng phân tích phương sai một nhân tố.

Bảng 6.4. Bảng phân tích phương sai một nhân tố thí nghiệm thiết kế theo kiểu CRD

Nguồn biến động	Tổng các bình phương (SS)	Độ tự do (df)	Trung bình bình phương (MS)	F tính toán	P
Nhân tố thí nghiệm	SS _B	t - 1	MS _B = SS _B /(t - 1)	MS _B /MS _W	
Sai số ngẫu nhiên	SS _W	N - t	MS _W = SS _W /(N - t)		
Tổng số	SS _T	N - 1			

Kết quả phân tích ANOVA ví dụ 4.3 được trình bày ở bảng 6.5.

Bảng 6.5. Bảng tổng hợp phân tích phương sai ảnh hưởng của mức protein trong khẩu phần đến tăng trọng của lợn ở ví dụ 4.3

Nguồn biến động	SS	df	MS	F tính toán	P
Mức protein	114,4375	2	57,21875	6,02	0,009
Sai số ngẫu nhiên	199,7188	21	9,510417		
Tổng số	314,1563	23			

Sai số chuẩn hay sai số của số trung bình (SEM) được ước tính theo công thức 6.2.

$$SEM = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{MS_W}{n}} \quad [6.2]$$

Khoảng tin cậy $100(1 - \alpha)$ cho trung bình của các nghiệm thức với cận trên cận dưới được tính theo công thức 6.3.

$$CI = M_i \pm t_{\alpha/2, (N-t)} SEM \quad [6.3]$$

Sai số của số trung bình về khả năng tăng trọng là:

$$SEM = \sqrt{\frac{MS_w}{n}} = \sqrt{\frac{9,510417}{8}} = 1,09$$

Khoảng tin cậy 95% của các trung bình là: $CI = M_i \pm t_{\alpha/2, (N-t)} SEM = M_i \pm 2,0796 \times 1,09 = M_i \pm 2,27$.

Nếu chúng ta bác bỏ giả thuyết H_0 thì chúng ta cần tiến hành phân tích *Post hoc* như đã đề cập ở phần Đa so sánh của chương 5. Có nhiều phương pháp để thực hiện đa so sánh, vấn đề cốt lõi là sử dụng phương pháp kiểm tra phù hợp. Kết quả phân tích ANOVA cho thấy $P < 0,05$ cho nên chúng ta cần thực hiện kiểm tra *Post hoc*. Cần lưu ý rằng mức ý nghĩa 0,05 là mức thường được các nhà khoa học chấp nhận, do vậy trong nội dung của giáo trình này chúng tôi sử dụng mức để quyết định chấp nhận hay bác bỏ H_0 là mức ý nghĩa $\alpha = 0,05$. Kết quả kiểm tra Tukey HSD được trình bày ở bảng 6.6. Giá trị Tukey HSD được tính theo công thức 4.44 (chương 4) là 3,58.

Bảng 6.6. Kết quả kiểm tra so sánh cặp theo phương pháp Tukey HSD đối với ví dụ 4.3

So sánh		$ M_i - M_j $	95% CI _{PC}		P
			Cận dưới	Cận trên	
13%CP	15%CP	4,25	0,35	8,15	0,031
13%CP	17%CP	4,87	0,97	8,78	0,013
15%CP	17%CP	0,62	- 3,28	4,53	0,914

Từ kết quả phân tích ANOVA và phân tích *Post hoc*, chúng ta kết luận:

1. Mức protein trong khẩu phần ảnh hưởng đến tăng trọng của lợn
2. Lợn ăn khẩu phần có mức protein 17% và 15%CP có tăng trọng cao hơn lợn ăn khẩu phần có mức protein 13%CP. Không có sự khác biệt về tăng trọng của lợn giữa khẩu phần có mức protein 15% và 17%CP.

6.2. Thiết kế thí nghiệm ngẫu nhiên theo khối đầy đủ

6.2.1. Định nghĩa

Kiểu thiết kế thí nghiệm ngẫu nhiên theo khối đầy đủ là kiểu thiết kế mà các đơn vị thí nghiệm được phân vào các khối, trong mỗi khối các đơn vị thí nghiệm được bố trí

một cách hoàn toàn ngẫu nhiên vào các nghiệm thức hoặc các nghiệm được bố trí một cách hoàn toàn ngẫu nhiên vào các đơn vị thí nghiệm; trong mỗi khối chứa đầy đủ các nghiệm thức.

6.2.2. Mô tả thiết kế thí nghiệm ngẫu nhiên theo khối đầy đủ

Kiểu thiết kế CRD không có sự khống chế phương sai của sai số ngẫu nhiên, do vậy nó chỉ có hiệu quả tốt khi các đơn vị thí nghiệm tương đồng. Trong thực tế không phải lúc nào chúng ta cũng có thể có được các đơn vị thí nghiệm tương đồng. Bất cứ nhân tố nào có ảnh hưởng đến biến phụ thuộc mà có sự khác nhau giữa các đơn vị thí nghiệm đều làm tăng phương sai của sai số ngẫu nhiên và đều làm giảm độ chính xác của thí nghiệm. Ví dụ các yếu tố như tuổi, lứa đẻ, khối lượng, giới tính đều là các yếu tố có thể ảnh hưởng đến các biến phụ thuộc ngoài yếu tố thí nghiệm. Để khống chế ảnh hưởng không mong muốn này chúng ta có thể nhóm các cá thể tương đồng vào một nhóm. Một nhóm như vậy gọi là một khối và chỉ tiêu để nhóm các đơn vị thí nghiệm gọi là yếu tố tạo khối. Ví dụ chúng ta nhóm vật nuôi theo khối lượng, theo tuổi, theo giai đoạn tiết sữa, hay theo lứa đẻ. Một ví dụ khác về các thí nghiệm trong phòng thí nghiệm, các thí nghiệm này sử dụng các điều chế viên như là yếu tố khối để xóa bỏ sự khác nhau về kết quả phân tích do các điều chế viên gây nên. Mỗi điều chế viên phân tích một lần lặp lại của các nghiệm thức.

Kiểu thiết kế thí nghiệm ngẫu nhiên theo khối đầy đủ ký hiệu là RCB (**R**andomized **C**omplete **B**lock Design). Khái niệm ngẫu nhiên ở đây có nghĩa là ngẫu nhiên trong mỗi khối. Khái niệm đầy đủ có nghĩa là trong mỗi khối chứa đầy đủ các nghiệm thức. RCB có thể được hiểu một cách đơn giản là một tập hợp các kiểu thiết kế theo CRD; mỗi khối tương đương với một CRD. RCB có một số ưu điểm sau:

- Nhờ vào sự phân các đơn vị thí nghiệm vào các nhóm tương đồng mà RCB đạt được kết quả chính xác hơn so với CRD khi các đơn vị thí nghiệm không tương đồng. Mức độ biến động giữa các đơn vị thí nghiệm càng lớn thì kiểu thiết kế RCB càng có hiệu quả hơn so với kiểu thiết kế CRD.
- RCB có thể linh động được số nghiệm thức và số lần lặp lại. Trong mỗi khối, mỗi nghiệm thức có thể chỉ có một lần lặp lại (trường hợp này thường xảy ra nhiều nhất) và có thể có số lần lặp lại là 2 hoặc có thể nhiều hơn.
- Phương pháp phân tích số liệu đơn giản. Khi số liệu của toàn bộ một khối bị mất hoặc của một hoặc hơn một nghiệm thức bị mất cũng không gây khó khăn gì cho phân tích.

Cần lưu ý rằng không phải trong mọi trường hợp thiết kế theo RCB luôn có hiệu quả hơn so với thiết kế theo CRD hay kiểu thiết kế thí nghiệm khác. Thiết kế theo kiểu RCB làm tiêu tốn độ tự do. Số độ tự do bị mất tương ứng với số khối - 1. Khi thay đổi độ tự do sẽ làm thay đổi giá trị MS_E và dẫn đến thay đổi giá trị F tính toán. Điều này có nghĩa là có thể thay đổi quyết định chấp nhận hay bác bỏ giả thuyết H_0 . Căn cứ vào hiệu

quả của các kiểu thiết kế để quyết định nên sử dụng kiểu thiết kế nào hiệu quả nhất. Hiệu quả của các kiểu thiết kế được tính toán dựa trên tỷ số các MS_E của các kiểu thiết kế. Trong nội dung của giáo trình này, chúng tôi không đề cập đến phương án tính hiệu quả của các kiểu thiết kế.

6.2.3. Ngẫu nhiên hóa

Phương pháp ngẫu nhiên hóa trong thiết kế RCB về cơ bản tuân theo nguyên tắc trình bày ở mục 6.2.3.2 “*Phương pháp ngẫu nhiên hóa*”. Điểm đặc biệt cần chú ý trong kiểu thiết kế này là ngẫu nhiên hóa trong mỗi khối. Quá trình ngẫu nhiên hóa trong một khối này không phụ thuộc vào quá trình ngẫu nhiên hóa trong khối khác. Số lần ngẫu nhiên hóa chính bằng số khối. Quá trình thiết kế thí nghiệm bắt đầu bởi việc sắp xếp các đơn vị thí nghiệm vào trong các nhóm tương đồng, sau đó tiến hành phân chia một cách ngẫu nhiên các nghiệm thức vào các đơn vị thí nghiệm trong mỗi nhóm tương đồng.

Ví dụ 6.1. Một nhóm quan tâm đến ảnh hưởng của 6 mức NSP (NSP1, NSP2, NSP3, NSP4, NSP5, NSP6) khác nhau trong khẩu phần đến hàm lượng Cresols trong phân lợn (mg/kg). Cơ sở thí nghiệm có 24 con lợn cùng giống, cùng giới tính, cùng tuổi tuy nhiên có khối lượng không đồng đều. Khối lượng là yếu tố có thể ảnh hưởng đến hàm lượng Cresols trong phân. Khối lượng ban đầu của lợn thí nghiệm được thể hiện qua bảng 6.7.

Bảng 6.7. Khối lượng ban đầu của lợn trong thí nghiệm ảnh hưởng của các mức NSP khác nhau trong khẩu phần đến hàm lượng Cresols trong phân lợn

68	42	33	51
42	37	68	72
51	33	54	52
42	57	35	56
50	46	62	48
54	36	42	44

Hãy tiến hành thí nghiệm và kiểm tra giả thuyết về ảnh hưởng của các mức NSP trong khẩu phần của lợn đến hàm lượng Cresols trong phân.

Do khối lượng lợn ban đầu có chênh lệch khá lớn cho nên chúng ta nghĩ ngay đến kiểu thiết kế RCB. Bước 1 của thiết kế theo RCB là sắp xếp các cá thể lợn theo khối lượng tăng dần hoặc giảm dần (cột 1, bảng 6.8). Bước tiếp theo là nhóm các cá thể có khối lượng tương đồng vào một khối. Trong ví dụ 6.1 chúng ta có 6 nghiệm thức và 24 cá thể lợn do vậy có thể có 4 khối, mỗi khối có 6 cá thể. Các cá thể trong mỗi khối có khối lượng giống nhau nhiều hơn so với các cá thể giữa các khối (cột 2, bảng 6.8). Bước tiếp theo là trong mỗi khối phân chia ngẫu nhiên các đơn vị thí nghiệm vào các nghiệm thức và kết quả được thể hiện ở cột 3 (bảng 6.8).

Bảng 6.8. Kết quả thiết kế thí nghiệm theo kiểu RCB của ví dụ 6.1

Cột 1 Khối lượng (kg)	Cột 2 Khối	Cột 3 Nghiệm thức
72	1	NSP6
68	1	NSP3
68	1	NSP4
62	1	NSP2
57	1	NSP5
56	1	NSP1
54	2	NSP3
54	2	NSP1
52	2	NSP4
51	2	NSP5
51	2	NSP6
50	2	NSP2
48	3	NSP3
46	3	NSP2
44	3	NSP5
42	3	NSP6
42	3	NSP4
42	3	NSP1
42	4	NSP6
37	4	NSP3
36	4	NSP2
35	4	NSP5
33	4	NSP1
33	4	NSP4

6.2.4. Mô hình thống kê và phân tích số liệu

6.2.4.1. Mô hình thống kê

Mô hình tuyến tính phân tích số liệu thí nghiệm thiết kế theo RCB có thể được hiểu một cách đơn giản từ mô hình phân tích số liệu thiết kế theo kiểu CRD. So với CRD, thiết kế thí nghiệm theo kiểu RCB có thêm một yếu tố cần khống chế là yếu tố khối, do vậy mô hình phân tích chính là mô hình trong thiết kế thí nghiệm theo kiểu CRD và thêm vào yếu tố khối.

$$y_{ij} = \mu + T_i + B_j + e_{ij} \quad [6.4]$$

Trong đó:

- y_{ij} là giá trị quan sát (trong ví dụ 6.1 là hàm lượng Cresols/kg phân) từ nghiệm thức i trong khối j . Trong mỗi khối, mỗi nghiệm thức được lặp lại một lần.

- μ là trung bình quần thể
- T_i là ảnh hưởng của nhân tố nghiên cứu (trong ví dụ 6.1 là mức NSP)
- B_j là ảnh hưởng của yếu tố khối, thể hiện sự sai lệch trung bình của các đơn vị thí nghiệm trong khối j so với trung bình tổng thể.
- e_{ij} là sai số thí nghiệm
- i là mức của nhân tố nghiên cứu (trong ví dụ 6.1, $i = 1, 2, \dots, t = 6$); j là số lần lặp lại $= n$ (trong ví dụ 6.1 $n = 4$)

Trong trường hợp thiết kế thí nghiệm RCB và mỗi nghiệm thức chỉ có một lần lặp lại thì ảnh hưởng của nghiệm thức và ảnh hưởng của khối được mặc định là cộng tính (không có sự tương tác giữa yếu tố khối và yếu tố thí nghiệm). Khi số lần lặp lại của mỗi nghiệm thức trong mỗi khối là 1 thì chúng ta không thể kiểm tra sự tương tác do không đủ độ tự do.

Để hiểu hơn về kiểu thiết kế thí nghiệm RCB cũng như mô hình phân tích chúng ta xem xét cách trình bày bảng số liệu trong kiểu thiết kế này ở bảng 6.9.

Bảng 6.9. Bảng trình bày số liệu trong thiết kế RCB (ví dụ 6.1)

Nghiệm thức	Khối				TB nghiệm thức
	1	2	3	4	
1 (NSP1)	y_{11}	y_{12}	y_{13}	y_{14}	M_1
2 (NSP2)	y_{21}	y_{22}	y_{23}	y_{24}	M_2
3 (NSP3)	y_{31}	y_{32}	y_{33}	y_{34}	M_3
4 (NSP4)	y_{41}	y_{42}	y_{43}	y_{44}	M_4
5 (NSP5)	y_{51}	y_{52}	y_{53}	y_{54}	M_5
6 (NSP6)	y_{61}	y_{62}	y_{63}	y_{64}	M_6
TB khối	B_1	B_2	B_3	B_4	GM

Sự chênh lệch của mỗi quan sát so với trung bình tổng thể được biểu diễn như sau:

$$y_{ij} - GM = (M_i - GM) + (B_j - GM) + (y_{ij} - M_i - B_j + GM) \quad [6.5]$$

Phía bên phải của mô hình gồm có 3 thành phần chính sau:

- Sai khác giữa ảnh hưởng của nghiệm thức với trung bình tổng thể: $M_i - GM$
- Sai khác giữa ảnh hưởng của khối với trung bình tổng thể: $B_j - GM$
- Sai số thí nghiệm: $y_{ij} - M_i - B_j + GM$

Tiếp tục biến đổi phương trình 6.5 ta có:

$$y_{ij} - M_i = (B_j - GM) + (y_{ij} - M_i - B_j + GM) \quad [6.6]$$

Sai số thí nghiệm $y_{ij} - M_i$ trong thiết kế thí nghiệm theo kiểu CRD được phân chia thành 2 phần: (1) do yếu tố khối gây ra $= B_j - GM$ và (2) sai số thí nghiệm trong kiểu thiết kế RCB $= y_{ij} - M_i - B_j + GM$. Tiếp tục biến đổi phương trình 6.7 ta có:

$$\sum_{i=1}^t \sum_{j=1}^n (y_{ij} - GM)^2 = n \sum_{i=1}^t (M_i - GM)^2 + t \sum_{j=1}^n (B_j - GM)^2 + \sum_{i=1}^t \sum_{j=1}^n (y_{ij} - M_i - B_j + GM)^2 \quad [6.7]$$

Hay

$$SS_{\text{tổng số}} = SS_{\text{nghiệm thức}} + SS_{\text{khối}} + SS_{\text{ngẫu nhiên}} \quad [6.8]$$

Như đã đề cập ở chương 4 điểm mấu chốt của phân tích phương sai là phân chia tổng biến động thành các biến động thành phần (6.6). Từ các biến động thành phần chúng ta có thể dễ dàng kiểm tra các giả thuyết nghiên cứu. Bảng tổng hợp kết quả phân tích phương sai trong kiểu thiết kế thí nghiệm RCB được trình bày ở bảng 6.10.

Bảng 6.10. Kết quả phân tích ANOVA trong thí nghiệm theo kiểu RCB

Nguồn biến động	SS	df	MS	F
Tổng số	$\sum_{i=1}^t \sum_{j=1}^n (y_{ij} - GM)^2$	nt - 1		
Nhân tố thí nghiệm	$n \sum_{i=1}^t (M_i - GM)^2$	t - 1	$MS_{\text{nghiệm thức}}$	$MS_{\text{nghiệm thức}}/MS_E$
Khối	$t \sum_{j=1}^n (B_j - GM)^2$	n - 1	$MS_{\text{khối}}$	
Phần còn lại	$\sum_{i=1}^t \sum_{j=1}^n (y_{ij} - M_i - B_j + GM)^2$	(t - 1)(n - 1)	MS_E	

6.2.4.2. Chuẩn bị số liệu để phân tích

Sau khi đã xác định rõ các thông tin như đơn vị thí nghiệm, số lần lặp lại, mô hình thống kê, chúng ta tiến hành thí nghiệm và thu thập số liệu. Khi đã có số liệu chúng ta cần chuẩn bị số liệu cho việc xử lý thống kê mô tả và thống kê suy diễn. Như đã trình bày ở trong phần thiết kế thí nghiệm theo kiểu CRD: nguyên tắc cơ bản nhất của chuẩn bị số liệu trong Excel là “mỗi hàng là một đơn vị thí nghiệm và mỗi cột là một nghiệm thức”. Đối với các kiểu thiết kế thí nghiệm có sự không chế ban đầu như RCB, LSD, Split - Plot hay Strip - Plot thì trong quá trình chuẩn bị số liệu, nhập số liệu và xử lý số liệu, yếu tố không chế (khối, hàng, cột, băng, dài) đều được xem như là yếu tố thí nghiệm.

**Bảng 6.11. Hàm lượng Cresols (mg/kg) trong phân lợn
được ăn bởi khẩu phần có hàm lượng NSP khác nhau**

Thí nghiệm	Khối			
	1	2	3	4
1 (NSP1)	34,98	41,22	36,94	39,97
2 (NSP2)	40,89	46,69	46,65	41,90
3 (NSP3)	42,07	49,42	52,68	42,91
4 (NSP4)	37,18	45,85	40,23	39,20
5 (NSP5)	37,99	41,99	37,61	40,45
6 (NSP6)	34,89	50,15	44,57	43,29

Kết quả phân tích hàm lượng Cresols (ví dụ 6.1) được trình bày ở bảng 6.11. Số liệu được chuẩn bị cho phân tích phương sai được trình bày ở bảng 6.12.

Bảng 6.12. Chuẩn bị số liệu cho phân tích số liệu của thí nghiệm thiết kế theo kiểu RCB

Khối lượng (kg)	Khối	Thí nghiệm	Cresols (mg/kg)
Cột 1	Cột 2	Cột 3	Cột 4
56	1	NSP1	34,98
62	1	NSP2	40,89
68	1	NSP3	42,07
68	1	NSP4	37,18
57	1	NSP5	37,99
72	1	NSP6	34,89
54	2	NSP1	41,22
50	2	NSP2	46,69
54	2	NSP3	49,42
52	2	NSP4	45,85
51	2	NSP5	41,99
51	2	NSP6	50,15
42	3	NSP1	36,94
46	3	NSP2	46,65
48	3	NSP3	52,68
42	3	NSP4	40,23
44	3	NSP5	37,61
42	3	NSP6	44,57
33	4	NSP1	39,97
36	4	NSP2	41,90
37	4	NSP3	42,91
33	4	NSP4	39,20
35	4	NSP5	40,45
42	4	NSP6	43,29

6.2.4.3. Phương pháp phân tích số liệu

Căn cứ vào mô hình thống kê để quyết định phương pháp phân tích số liệu. Chúng ta cần phân tích hai loại thống kê cơ bản: thống kê mô tả và phân tích thống kê suy diễn. Chúng ta chỉ quan tâm đến các giá trị thống kê mô tả về hàm lượng Cresols trong phân theo các mức NSP trong khẩu phần. Thông thường chúng ta không quan tâm đến các tham số thống kê mô tả cũng như sự khác nhau về hàm lượng Cresols giữa các khối. Nói cách khác, khối không phải là nhân tố nghiên cứu.

Kết quả phân tích thống kê mô tả về hàm lượng Cresols (ví dụ 6.1) được thể hiện ở bảng 6.13. Để có thể kiểm tra giả thuyết về ảnh hưởng của các mức NSP trong khẩu phần đến hàm lượng Cresols trong phân lợn chúng ta cần tiến hành phân tích phương sai. Kết quả phân tích phương sai được thể hiện ở bảng 6.14. Qua bảng 6.14 ta thấy rằng hàm lượng NSP trong khẩu phần thức ăn có ảnh hưởng đến hàm lượng Cresols trong phân lợn ($F_{\text{tính toán}} = 5,592 > F_{\text{tra bảng}}$ ở mức 0,05 và độ tự do 5 và 15 = 2,90). Ở bảng 6.14 chúng ta cũng thấy yếu tố khối cũng có ảnh hưởng đến hàm lượng Cresols. Điều cần lưu ý là chúng ta không cần quan tâm đến ảnh hưởng của yếu tố khối. Chúng ta không ngạc nhiên khi thấy rằng yếu tố khối có ảnh hưởng đến biến nghiên cứu, đây cũng chính là lý do chúng ta áp dụng kiểu thiết kế RCB. Một số phần mềm chuyên dụng như Genstat thì kết quả về ảnh hưởng của yếu tố khối thường không được đề cập. Giá trị F tính toán cho yếu tố khối cũng không được trình bày. Tuy nhiên, một số phần mềm khác như SPSS thì giá trị F tính toán, mức độ tin cậy của yếu tố khối được trình bày trong bảng phân tích phương sai.

Bảng 6.13. Một số tham số thống kê mô tả về hàm lượng Cresols trong phân lợn (mg/kg) ở ví dụ 6.1

Tham số	NSP1	NSP2	NSP3	NSP4	NSP5	NSP6
Trung bình	38,28	44,03	46,77	40,62	39,51	43,23
Sai số của số TB	1,42	1,54	2,56	1,86	1,04	3,15
Trung vị	38,46	44,28	46,17	39,72	39,22	43,93
Độ lệch chuẩn	2,84	3,07	5,13	3,71	2,08	6,30
Phương sai	8,06	9,45	26,31	13,78	4,32	39,75
Khoảng biến động	6,24	5,80	10,61	8,67	4,38	15,26

Bảng 6.14. Bảng phân tích phương sai một nhân tố thí nghiệm thiết kế theo kiểu RCB

Nguồn biến động	SS	df	MS	$F_{\text{tính toán}}$	P
NSP	201,316	5	40,263	5,592	0,004
Khối	197,004	3	65,668	9,120	0,001
Ngẫu nhiên	108,008	15	7,201		
Tổng số	506,329	23			

Sai số chuẩn hay sai số của số trung bình (SEM) được ước tính theo công thức 6.9.

$$SEM = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{MS_W}{n}} = \sqrt{\frac{7,2}{4}} = 1,34 \quad [6.9]$$

Khoảng tin cậy $100(1 - \alpha)$ cho trung bình của các nghiệm thức với cận trên cận dưới được tính theo công thức 6.10.

$$CI = M_i \pm t_{\alpha/2, (N-1)} SEM = M_i \pm 2,131 \times 1,34 = M_i \pm 2,86 \quad [6.10]$$

Sai số chuẩn của sự khác nhau giữa 2 nghiệm thức (SED) được tính theo công thức 6.11.

$$SED = \sqrt{\frac{2S^2}{n}} = \sqrt{\frac{2MS_W}{n}} = \sqrt{\frac{2 \times 7,2}{4}} = 1,825 \quad [6.11]$$

Nếu chúng ta bác bỏ giả thuyết H_0 thì chúng ta cần tiến hành phân tích *Post hoc* như đã đề cập ở phần Đa so sánh của chương 4. Có nhiều phương pháp để thực hiện đa so sánh, vấn đề cốt lõi là sử dụng phương pháp kiểm tra phù hợp. Kết quả phân tích ANOVA (bảng 6.14) cho thấy $P = 0,004 < 0,05$ cho nên chúng ta cần thực hiện kiểm tra *Post hoc*.

Kết quả kiểm tra Tukey HSD được trình bày ở bảng 6.15. Giá trị Tukey HSD được tính theo công thức 4.44 là 6,16.

Bảng 6.15. Kết quả kiểm tra so sánh cặp theo phương pháp Tukey HSD đối với ví dụ 6.1

So sánh		$ M_i - M_j $	P	95% CI _{PC}	
				Cận dưới	Cận trên
NSP2	NSP1	5,76	0,07	-0,41	11,92
NSP2	NSP4	3,42	0,49	-2,75	9,58
NSP2	NSP5	4,52	0,22	-1,64	10,69
NSP2	NSP6	0,81	1,00	-5,36	6,97
NSP3	NSP1	8,49	0,00	2,33	14,66
NSP3	NSP2	2,74	0,70	-3,43	8,90
NSP3	NSP4	6,16	0,05	-0,01	12,32
NSP3	NSP5	7,26	0,02	1,10	13,42
NSP3	NSP6	3,55	0,46	-2,62	9,71
NSP4	NSP1	2,34	0,81	-3,83	8,50
NSP4	NSP5	1,11	0,99	-5,06	7,27
NSP5	NSP1	1,23	0,98	-4,93	7,40
NSP6	NSP1	4,95	0,16	-1,22	11,11
NSP6	NSP4	2,61	0,74	-3,55	8,77
NSP6	NSP5	3,72	0,41	-2,45	9,88

Từ kết quả phân tích ANOVA và HSD, chúng ta có thể kết luận:

1. Hàm lượng NSP trong khẩu phần có ảnh hưởng đến hàm lượng Cresols trong phân của lợn
2. Lợn được ăn khẩu phần với mức NSP3 cho hàm lượng Cresols trong phân cao hơn mức NSP1, NSP4 và NSP5.

6.2.4.4. Nhiều đơn vị thí nghiệm cho mỗi nghiệm thức trong mỗi khối

Trong ví dụ 6.1 ở trên ta có mỗi nghiệm thức được lặp lại một lần trong mỗi khối. Trong thực tế nhiều khi chúng ta có hơn một lần lặp lại cho mỗi nghiệm thức trong mỗi khối. Nếu trường hợp này xảy ra thì chúng ta có thể kiểm tra sự tương tác giữa nhân tố thí nghiệm và yếu tố khối (chúng ta có đủ độ tự do để kiểm tra sự tương tác). Khi đó mô hình phân tích là:

$$y_{ijk} = \mu + T_i + B_j + (TB)_{ij} + e_{ijk} \quad [6.12]$$

Trong đó:

- $i = 1, 2, \dots, t$; $j = 1, 2, \dots, b$; $k = 1, 2, \dots, u$
- T_i là ảnh hưởng của nhân tố thí nghiệm
- B_j là ảnh hưởng của khối
- $(TB)_{ij}$ là ảnh hưởng của tương tác giữa khối và nhân tố thí nghiệm
- e_{ijk} là sai số ngẫu nhiên

Nói một cách tổng quát từ mô hình trên: thí nghiệm có t nghiệm thức, b khối và trong mỗi khối mỗi đơn vị thí nghiệm được lặp lại u lần.

Bảng 6.16. Bảng trình bày số liệu trong thiết kế RCB có hơn một lần lặp lại cho mỗi nghiệm thức trong mỗi khối

Nghiệm thức	Khối				Trung bình nghiệm thức
	1	2	3	4	
1 (NSP1)	y_{111}	y_{121}	y_{131}	y_{141}	M_1
	y_{112}	y_{122}	y_{132}	y_{142}	
2 (NSP2)	y_{211}	y_{221}	y_{231}	y_{241}	M_2
	y_{212}	y_{222}	y_{232}	y_{242}	
3 (NSP3)	y_{311}	y_{321}	y_{331}	y_{341}	M_3
	y_{312}	y_{322}	y_{332}	y_{342}	
4 (NSP4)	y_{411}	y_{421}	y_{431}	y_{441}	M_4
	y_{412}	y_{422}	y_{432}	y_{442}	
5 (NSP5)	y_{511}	y_{521}	y_{531}	y_{541}	M_5
	y_{512}	y_{522}	y_{532}	y_{542}	
6 (NSP6)	y_{611}	y_{621}	y_{631}	y_{641}	M_6
	y_{612}	y_{622}	y_{632}	y_{642}	
Trung bình khối	B_1	B_2	B_3	B_4	GM

Quay trở lại ví dụ 6.1, đặt giả sử có 2 lần lặp lại cho mỗi nghiệm thức trong mỗi khối. Như vậy, thí nghiệm có $t = 6$ nghiệm thức, $b = 4$ khối và $u = 2$ lần lặp lại cho mỗi nghiệm thức trong mỗi khối. Khi đó bảng trình bày số liệu thí nghiệm được thể hiện ở bảng 6.16.

Như đã trình bày ở chương 4, mẫu chốt của phân tích phương sai là phân chia tổng biến động thành các biến động thành phần. Chúng ta có các thành phần biến động như sau:

$$SS_{\text{tổng số}} = \sum_{i=1}^t \sum_{j=1}^b \sum_{k=1}^u (y_{ijk} - GM)^2 \quad [6.13]$$

$$SS_{\text{nghiệm thức}} = bu \sum_{i=1}^t (M_i - GM)^2 = \sum_{i=1}^t \sum_{j=1}^b \sum_{k=1}^u (M_i - GM)^2 \quad [6.14]$$

$$SS_{\text{khối}} = tu \sum_{j=1}^b (B_j - GM)^2 = \sum_{i=1}^t \sum_{j=1}^b \sum_{k=1}^u (B_j - GM)^2 \quad [6.15]$$

$$SS_{\text{tương tác}} = u \sum_{i=1}^t \sum_{j=1}^b (TB_{ij} - GM)^2 - SS_{\text{khối}} - SS_{\text{nghiệm thức}} \quad [6.16]$$

$$SS_E = SS_{\text{tổng số}} - SS_{\text{nghiệm thức}} - SS_{\text{khối}} - SS_{\text{tương tác}} \quad [6.17]$$

Từ các biến động thành phần chúng ta có thể dễ dàng tính được các trung bình bình phương và từ đó tính toán được các giá trị F tính toán để kiểm tra các giả thuyết H_0 . Bảng tổng hợp kết quả phân tích phương sai trong kiểu thiết kế thí nghiệm RCB có hơn một lần lặp lại cho mỗi nghiệm thức trong mỗi khối được trình bày ở bảng 6.17.

Bảng 6.17. Kết quả ANOVA trong thí nghiệm theo kiểu RCB có hơn một lần lặp lại cho mỗi nghiệm thức trong mỗi khối

Nguồn biến động	SS	df	MS	F _{tính toán}
Tổng số	$\sum_{i=1}^t \sum_{j=1}^b \sum_{k=1}^u (y_{ijk} - GM)^2$	$tbu - 1$		
Nhân tố thí nghiệm	$bu \sum_{i=1}^t (M_i - GM)^2 = \sum_{i=1}^t \sum_{j=1}^b \sum_{k=1}^u (M_i - GM)^2$	$t - 1$	$MS_{\text{nghiệm thức}}$	$MS_{\text{nghiệm thức}} / MS_E$
Khối	$tu \sum_{j=1}^b (B_j - GM)^2 = \sum_{i=1}^t \sum_{j=1}^b \sum_{k=1}^u (B_j - GM)^2$	$b - 1$	$MS_{\text{khối}}$	
Nhân tố × Khối	$\sum_{i=1}^t \sum_{j=1}^b (TB_{ij} - GM)^2 - SS_{\text{khối}} - SS_{\text{nghiệm thức}}$	$(t - 1)(b - 1)$	$MS_{\text{khối} \times \text{nhân tố}}$	$MS_{\text{khối} \times \text{nhân tố}} / MS_E$
Sai số ngẫu nhiên	$SS_E = SS_{\text{tổng số}} - SS_{\text{nghiệm thức}} - SS_{\text{khối}} - SS_{\text{tương tác}}$	$tb(u - 1)$	MS_E	

Từ giá trị F tính toán ta so sánh với giá trị F tra bảng để quyết định chấp nhận hay bác bỏ giả thuyết H_0 . Giá trị F tra bảng của nhân tố thí nghiệm là $F_{\alpha, (t-1), tb(u-1)}$. Giá trị F tra bảng của sự tương tác giữa nhân tố và khối là $F_{\alpha, (b-1), tb(u-1)}$.

6.2.4.5. Thiếu số liệu trong thiết kế thí nghiệm theo RCB

Nhiều khi trong quá trình làm thí nghiệm thiết kế theo kiểu RCB chúng ta bị mất đi một số liệu nào đó. Ví dụ một con lợn nào đó bị chết hay một con gà nào đó bị chuột ăn... Trong trường hợp như vậy chúng ta có thể bỏ toàn bộ khối số liệu mà có số liệu bị mất. Tuy nhiên việc bỏ số liệu toàn bộ khối như vậy dẫn đến mất thông tin. Li (1964) đã đề xuất phương pháp điều chỉnh số phân tích phương sai trong trường hợp có số liệu bị mất. Giá trị của số liệu bị mất thuộc nghiệm thức i trong khối j có thể được tính như sau:

$$\hat{y}_{ij} = \frac{tM_i + bB_j - \sum_{i=1}^t \sum_{j=1}^b y_{ij}}{(t-1)(b-1)} \quad [6.18]$$

Trong đó, M_i là tổng của các số liệu đã biết của nghiệm thức i và B_j là tổng các số liệu đã biết trong khối j và $\sum_{i=1}^t \sum_{j=1}^b y_{ij}$ là tổng của các số liệu đã biết. Tổng bình phương tổng số, nhân tố thí nghiệm, khối được tính theo trường hợp mất số liệu như đã đề cập ở trên, ngoại trừ SS nghiệm thức bị chênh lệch một chút, để giải quyết hạn chế này ta trừ đi một lượng sau đây:

$$\text{Lượng chênh lệch} = \frac{[B_i - (t-1)\hat{y}_{ij}]^2}{t(t-1)} \quad [6.19]$$

và df_T và df_E trừ đi 1.

6.3. Thiết kế thí nghiệm hình vuông la tinh

6.3.1. Định nghĩa

Thiết kế thí nghiệm hình vuông la tinh là kiểu thiết kế mà các nghiệm thức được phân vào các đơn vị thí nghiệm theo hai hướng khác nhau (hàng và cột).

6.3.2. Mô tả thiết kế thí nghiệm hình vuông la tinh

Kiểu thiết kế thí nghiệm hình vuông la tinh hay ô vuông la tinh được ký hiệu là LSD (Latin Square Design). Thuật ngữ hình vuông la tinh xuất phát từ đế chế La mã cổ đại. Vía hè thời La Mã cổ đại thường được lát đá hoa có màu khác nhau ở các hàng và cột. Mỗi màu chỉ xuất hiện một lần duy nhất trong mỗi hàng và mỗi cột. Triết lý của thiết kế thí nghiệm LSD được mô phỏng từ đó.

Một trong ba nguyên tắc cơ bản của thiết kế thí nghiệm là khống chế sự sai khác ban đầu. Thiết kế thí nghiệm RCB là một cách để khống chế sự sai khác ban đầu. Tuy nhiên RCB chỉ khống chế được một yếu tố (ngoài yếu tố nghiên cứu) có thể ảnh hưởng đến biến phụ thuộc. Trong thực tế nhiều khi có nhiều hơn một biến động cần phải được khống chế để tăng độ chính xác của thí nghiệm. Điều này nảy sinh thí nghiệm theo kiểu LSD, có khả năng khống chế hơn một biến động ví dụ các cá thể thí nghiệm có sự khác nhau về lứa đẻ và khối lượng ban đầu, các cá thể bò khác nhau ở các giai đoạn tiết sữa khác nhau.

Đặc trưng của kiểu thiết kế LSD thuần túy là số hàng (R) bằng số cột (C) và bằng số nghiệm thức (t). Số đơn vị thí nghiệm chính là bình phương của số nghiệm thức, hay số hàng, hay số cột. Đơn vị thí nghiệm trong kiểu thiết kế thí nghiệm này không phải là các cá thể vật nuôi hay nhóm cá thể vật nuôi nghiên cứu mà là sự kết hợp giữa yếu tố hàng và yếu tố cột, ví dụ sự kết hợp giữa cá thể và giai đoạn.

Kiểu thiết kế thí nghiệm này thường có hiệu quả đối với thí nghiệm có số lượng động vật ít. Ví dụ, để đánh giá ảnh hưởng của 4 loại khẩu phần khác nhau đến tỷ lệ tiêu hóa các chất, chúng ta có thể thiết kế thí nghiệm theo LSD với 4 cá thể thí nghiệm. Trong khi đó cũng câu hỏi nghiên cứu này nếu thiết kế theo kiểu CRD hoặc RCB thì cần số lượng động vật thí nghiệm lớn hơn rất nhiều so với LSD.

Kiểu thiết kế thí nghiệm này kiểm soát được biến động theo 2 hướng (hàng và cột), do vậy về mặt cơ bản kiểu thiết kế thí nghiệm này có hiệu quả cao hơn so với kiểu thiết kế CRD và RCB. Có thể hiểu một cách đơn giản nhất là hàng và cột trong thiết kế thí nghiệm LSD chính là các thiết kế thí nghiệm RCB.

Một điểm lưu ý trong kiểu thiết kế thí nghiệm này là khoảng thời gian nghỉ giữa 2 lần thí nghiệm. Mục tiêu là xóa bỏ ảnh hưởng kéo dài từ nghiệm thức trước ra nghiệm thức sau. Để đạt được điều này có thể thực hiện được khi có thời gian nghỉ thích ứng. Kiểu thiết kế thí nghiệm này không kiểm tra sự tương tác giữa yếu tố hàng và yếu tố cột và giả định thông thường bị vi phạm nếu chúng ta không lựa chọn biến độc lập đúng và không có thời gian nghỉ giữa 2 giai đoạn kế tiếp nhau hợp lý.

Kiểu thiết kế thí nghiệm này không nên tiến hành khi biến độc lập có ảnh hưởng lâu dài, ảnh hưởng tích lũy đến biến phụ thuộc. Bảng 6.18 tóm tắt một số trường hợp thích hợp và không thích hợp với kiểu thiết kế LSD.

Bảng 6.18. Một số trường hợp nghiên cứu thích hợp và không thích hợp cho kiểu thiết kế thí nghiệm hình vuông la tinh

<i>Có thể sử dụng kiểu ô vuông la tinh</i>	<i>Không thích hợp cho kiểu ô vuông la tinh</i>
- Ảnh hưởng của khẩu phần đến thành phần trong sữa	- Ảnh hưởng của năng lượng hoặc protein ăn vào đến năng suất sữa
- Ảnh hưởng của khẩu phần đến khối lượng trứng nnnnnnn	- Ảnh hưởng của năng lượng và protein ăn vào đến tốc độ đẻ trứng
- Ảnh hưởng của xử lý hormon	- Ảnh hưởng của bất cứ nhân tố nào đến sinh trưởng
- Thí nghiệm tiêu hóa	- Ảnh hưởng của các biện pháp quản lý đến bệnh viêm vú

Để thiết lập một bố trí thí nghiệm LSD, số lượng cá thể và số giai đoạn phải tương đương với số nghiệm thức và điều này nhìn chung hạn chế tính ứng dụng của nó trong trường hợp có 4 đến 5 nghiệm thức cần so sánh. Ô vuông 2×2 không có độ tự do của ngẫu nhiên, do vậy không thể tiến hành suy diễn thống kê nên chúng ta không thực hiện thí nghiệm thiết kế theo kiểu LSD 2×2 . Ô vuông 3×3 chỉ có 2 df ngẫu nhiên, như vậy chúng ta cần phải có đủ df để bảo đảm độ chính xác. Ô vuông lớn hơn 5×5 thì lại rất tốn thời gian để tiến hành thí nghiệm, do đó nó thường không mang tính thực tế. Trong thực tế thiết kế thí nghiệm LSD 4×4 thường được sử dụng.

6.3.3. Phương pháp ngẫu nhiên hóa trong thiết kế thí nghiệm kiểu ô vuông la tinh

Ví dụ 6.2. Một nhóm nghiên cứu ảnh hưởng của 4 loại thức ăn bổ sung khác nhau (A, B, C, D) đến năng suất sữa của bò. Cơ sở thí nghiệm có 4 con bò đang ở cùng giai đoạn tiết sữa. Hãy thiết kế thí nghiệm và nêu phương án phân tích số liệu. Giả sử bò đang ở giai đoạn tiết sữa ổn định.

Để trả lời câu hỏi nghiên cứu trên, nhóm nghiên cứu có thể thiết kế thí nghiệm theo kiểu LSD. Tiến hành ngẫu nhiên hóa trong kiểu thiết kế thí nghiệm LSD tùy thuộc việc chúng ta đã có hình vuông la tinh chuẩn hay chưa.

Bước 1: Lấy hình vuông la tinh chuẩn. Tất cả các hình vuông la tinh với một kích thước xác định đều có thể lấy từ các hình vuông la tinh chuẩn. Hình vuông la tinh chuẩn có các ký hiệu của nghiệm thức là A, B, C, D và các ký hiệu này được viết theo thứ tự alphabet trong dòng đầu tiên và hàng đầu tiên (biểu đồ 6.1).

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Biểu đồ 6.1. Sơ đồ một hình vuông la tinh chuẩn 4×4

Bước 2: Ngẫu nhiên hóa trật tự tất cả các hàng ngoại trừ hàng thứ nhất. Chú ý, nếu chúng ta không có hình vuông la tinh chuẩn thì ở bước 2 này chúng ta nên ngẫu nhiên hóa tất cả các hàng.

Để thực hiện bước 2 ta cần có các số ngẫu nhiên cho 3 hàng cuối (có thể dùng RAND(), có thể dùng bảng các số ngẫu nhiên) để sắp xếp trật tự 3 hàng cuối.

Số ngẫu nhiên	Hàng gốc
3	2
1	3
2	4

Như vậy, trật tự mới của các nghiệm thức trong biểu đồ 6.1 được thể hiện ở biểu đồ 6.2.

A	B	C	D
C	D	A	B
D	A	B	C
B	C	D	A

Biểu đồ 6.2. Trật tự các nghiệm thức sau khi đã được ngẫu nhiên hóa trật tự của các hàng

Bước 3: Ngẫu nhiên hóa trật tự của cột. Cần có các số ngẫu nhiên để sắp xếp trật tự các cột.

Số ngẫu nhiên	Cột gốc
1	1
4	2
3	3
2	4

Trật tự mới của các nghiệm thức ở biểu đồ 6.2 được thể hiện ở bảng 6.3.

1	4	3	2
A	D	C	B
C	B	A	D
D	C	B	A
B	A	D	C

Biểu đồ 6.3. Trật tự các nghiệm thức sau khi đã được ngẫu nhiên hóa trật tự của các hàng và các cột

Bước 4: Phân chia một cách ngẫu nhiên các nghiệm thức vào các ký tự trong hình vuông la tinh chuẩn.

Để có thể tiến hành bước 4 thì cần có các số ngẫu nhiên để phân các nghiệm thức vào các ký tự. Nếu hình vuông la tinh chuẩn đã được chọn một cách ngẫu nhiên từ các hình vuông la tinh chuẩn có khả năng thì bước này không cần thiết.

Giả sử trong trường hợp này chúng ta đã không chọn một cách ngẫu nhiên hình vuông la tinh chuẩn từ các hình vuông la tinh chuẩn có khả năng. Khi đó chúng ta cần thực hiện bước 4. Giả sử các nghiệm thức là W, X, Y và Z. Các nghiệm thức W, X, Y và Z thay các ký tự trong hình vuông la tinh theo thứ tự D, B, C và A theo một bố trí ngẫu nhiên.

Số ngẫu nhiên	Ký tự	Nghiệm thức
4	D	W
2	B	X
3	C	Y
1	A	Z

6.3.4. Mô hình thống kê và phân tích số liệu

6.3.4.1. Mô hình thống kê

Mô hình tuyến tính phân tích số liệu thí nghiệm thiết kế theo LSD có thể được hiểu một cách đơn giản từ mô hình phân tích số liệu thiết kế theo kiểu RCB. So với RCB, thiết kế thí nghiệm theo kiểu LSD có thêm một yếu tố cần khống chế là yếu tố hàng hoặc cột, do vậy mô hình phân tích chính là mô hình trong thiết kế thí nghiệm theo kiểu RCB và thêm vào một yếu tố cần khống chế.

$$y_{ij} = \mu + R_i + C_j + T_k + e_{ij} \quad [6.20]$$

với $i, j, k = 1, 2, \dots, t$

Trong đó:

- y_{ij} là giá trị quan sát (trong ví dụ 6.2 là năng suất sữa) của đơn vị thí nghiệm trong hàng j và cột k .
- μ là trung bình quần thể
- T_k là ảnh hưởng của nghiệm thức thứ i của nhân tố nghiên cứu (trong ví dụ 6.2 là loại thức ăn bổ sung)
- R_i là ảnh hưởng của yếu tố hàng
- C_j là ảnh hưởng của yếu tố cột
- e_{ij} là sai số thí nghiệm

Phương pháp phân tích một thiết kế thí nghiệm theo kiểu LSD tương tự như phương pháp phân tích đối với thiết kế thí nghiệm theo khối ngẫu nhiên đầy đủ mà có thêm một nguồn biến động. Thay vì đề cập đến các khối trong thiết kế thí nghiệm theo khối thì trong thiết kế thí nghiệm theo kiểu LSD đề cập đến hai hình thức biến động đó là *hàng* và *cột*. Như vậy, giá trị của một tính trạng = ảnh hưởng của hàng + ảnh hưởng của cột + ảnh hưởng của nghiệm thức + ảnh hưởng của ngẫu nhiên. Tương tự như vậy, chúng ta có thể chia tổng biến động thành biến động của hàng, cột, nghiệm thức và ngẫu nhiên.

Bảng 6.19. Kết quả ANOVA trong thí nghiệm theo kiểu LSD

Nguồn biến động	df	Tổng biến động	MS	F
Hàng	t - 1	$\frac{R_1^2 + R_2^2 + \dots + R_t^2}{t} - \frac{\left(\sum_{i=1}^t x\right)^2}{t^2}$ Hay $t \sum_i (M_i - GM)^2$	MS _R	
Cột	t - 1	$\frac{C_1^2 + C_2^2 + \dots + C_t^2}{t} - \frac{\left(\sum_{j=1}^t x\right)^2}{t^2}$ Hay $t \sum_j (M_j - GM)^2$	MS _C	
Nghiem thức	t - 1	$\frac{T_1^2 + T_2^2 + \dots + T_t^2}{t} - \frac{\left(\sum_{k=1}^t x\right)^2}{t^2}$ Hay $t \sum_k (M_k - GM)^2$	MS _T	$\frac{MS_T}{MS_E}$
Ngẫu nhiên	(t - 1)(t - 2)	Bằng cách trừ SS _{tổng số} cho các tổng các bình phương độ lệch khác	MS _E	
Tổng số	t ² - 1	$\sum_{i=1}^{t^2} x^2 - \frac{\left(\sum_{i=1}^{t^2} x\right)^2}{t^2}$ Hay $\sum_i \sum_j (M_{ij} - GM)^2$		

Như vậy, SS_E bị giảm đi so với SS_E trong thiết kế thí nghiệm theo kiểu RCB một lượng tương đương với SS_{hàng} hoặc SS_{cột}. Đồng thời với bị giảm SS_E thì độ tự do của ngẫu nhiên cũng bị mất đi t - 1. Như vậy, lực thống kê sẽ bị giảm nếu như việc giảm SS_E do yếu tố hàng và cột là không lớn.

6.3.4.2. Chuẩn bị số liệu để phân tích

Sau khi đã xác định rõ các thông tin như đơn vị thí nghiệm, số lần lặp lại, mô hình thống kê, chúng ta tiến hành thí nghiệm và thu thập số liệu. Khi đã có số liệu chúng ta

cần chuẩn bị số liệu cho việc xử lý thống kê mô tả và thống kê suy diễn. Như đã trình bày ở trong phần thiết kế thí nghiệm theo kiểu CRD: nguyên tắc cơ bản nhất của chuẩn bị số liệu trong Excel là “mỗi hàng là một đơn vị thí nghiệm và mỗi cột là một nghiệm thức”. Đối với các kiểu thiết kế thí nghiệm có sự khống chế ban đầu như RCB, LSD, Split - Plot hay Strip - Plot thì trong quá trình chuẩn bị số liệu, nhập số liệu và xử lý số liệu, yếu tố khống chế đều được xem như là yếu tố thí nghiệm. Giả sử kết quả nghiên cứu ảnh hưởng của 4 loại thức ăn bổ sung đến năng suất sữa (ví dụ 6.2) được thể hiện ở bảng 6.20.

Bảng 6.20. Năng suất sữa của 4 con bò được ăn bởi 4 loại khẩu phần khác nhau

Giai đoạn	Cá thể 1	Cá thể 2	Cá thể 3	Cá thể 4
1	A 192	D 203	C 292	B 163
2	C 214	B 195	A 218	D 249
3	D 221	C 152	B 204	A 134
4	B 190	A 139	D 245	C 210

Các số liệu được nhập như sau (bảng 6.21):

Bảng 6.21. Chuẩn bị số liệu cho phân tích số liệu của thí nghiệm thiết kế theo kiểu LSD

Giai đoạn	Cá thể	Khẩu phần	Năng suất (pound/tuần)
I	1	A	192
I	2	D	203
I	3	C	292
I	4	B	163
II	1	C	214
II	2	B	195
II	3	A	218
II	4	D	249
III	1	D	221
III	2	C	152
III	3	B	204
III	4	A	134
IV	1	B	190
IV	2	A	139
IV	3	D	245
IV	4	C	210

6.2.4.3. Phương pháp phân tích số liệu

Kết quả phân tích mô tả được thể hiện qua bảng 6.22.

Bảng 6.22. Một số tham số thống kê mô tả về sản lượng sữa bò ở các loại thức ăn bổ sung khác nhau (pound/tuần) ở ví dụ 6.2

Tham số	Khẩu phần A	Khẩu phần B	Khẩu phần C	Khẩu phần D
Trung bình	170,8	188,0	217,0	229,5
Sai số của số trung bình	20,5	8,8	28,7	10,8
Trung vị	165,5	192,5	212,0	233,0
Độ lệch chuẩn	41,0	17,6	57,5	21,6
Phương sai	1680,9	311,3	3302,7	465,0
Khoảng biến động	84,0	41,0	140,0	46,0
Số mẫu	4,0	4,0	4,0	4,0

Kết quả phân tích phương sai được thể hiện ở bảng 6.23.

Bảng 6.23. Kết quả phân tích phương sai về ảnh hưởng của các loại thức ăn bổ sung đến năng suất sữa của bò ở ví dụ 6.2

Nguồn biến động	SS	df	MS	F	P
Cả thể	9929,2	3	3309,73	6,08	0,03
Giai đoạn	4085,7	3	1361,90	2,50	0,16
Khẩu phần	8607,7	3	2869,23	5,27	0,04
Ngẫu nhiên	3264,9	6	544,15		
Tổng số	25887,4	15			

Kết quả ANOVA ở bảng 6.23 cho thấy chúng ta bác bỏ H_0 ở mức $\alpha = 0,05$ ($P < 0,05$ hay giá trị $F_{\text{tính toán}} = 5,27 > \text{giá trị } F_{\text{tra bảng}}(0,05; 3; 15) = 3,2873$). Điều này có nghĩa các loại thức ăn bổ sung có ảnh hưởng đến năng suất sữa.

Sai số chuẩn hay sai số của số trung bình (SEM) được ước tính theo công thức 6.21.

$$SEM = \sqrt{\frac{S^2}{t}} = \sqrt{\frac{MS_w}{t}} = \sqrt{\frac{544,15}{4}} = 11,66 \quad [6.22]$$

Khoảng tin cậy $100(1 - \alpha)$ cho trung bình của các nghiệm thức với cận trên cận dưới được tính theo công thức 6.23.

$$CI = M_i \pm t_{\alpha/2, (N-t)} SEM = M_i \pm 2,1314 \times 11,66 = M_i \pm 24,85 \quad [6.23]$$

Sai số chuẩn của sự khác nhau giữa 2 nghiệm thức (SED) được tính theo công thức 6.24.

$$SED = \sqrt{\frac{2S^2}{t}} = \sqrt{\frac{2MS_w}{t}} = \sqrt{\frac{2 \times 544,15}{4}} = 16,50 \quad [6.24]$$

Bảng 6.24. Kết quả kiểm tra so sánh cặp theo phương pháp Tukey HSD đối với ví dụ 6.2

So sánh		$M_i - M_j$	P	95% CI	
				Cận dưới	Cận trên
B	A	17,25	0,731	- 39,85	74,35
C	A	46,25	0,109	- 10,85	103,35
C	B	29,00	0,375	- 28,10	86,10
D	A	58,75	0,045	1,65	115,85
D	B	41,50	0,154	- 15,60	98,60
D	C	12,50	0,870	- 44,60	69,60

Nếu chúng ta bác bỏ giả thuyết H_0 thì chúng ta cần tiến hành phân tích *Post hoc* hoặc so sánh có kế hoạch như đã đề cập ở phần Đa so sánh của chương 4. Có nhiều phương pháp để thực hiện đa so sánh, vấn đề cốt lõi là sử dụng phương pháp kiểm tra phù hợp. Kết quả phân tích ANOVA (bảng 6.23) cho thấy $P = 0,04 < 0,05$ cho nên chúng ta cần thực hiện kiểm tra *Post hoc* hoặc so sánh có kế hoạch.

Giả sử chúng ta quan tâm đến sự khác nhau giữa các nghiệm thức về ảnh hưởng của nó đến năng suất sữa, chúng ta có tiến hành kiểm tra Tukey HSD. Kết quả kiểm tra Tukey HSD được trình bày ở bảng 6.24. Như vậy thức ăn bổ sung D đưa lại năng suất sữa cao hơn thức ăn bổ sung A ở mức $P = 0,045$.

6.3.5. Thiết kế thí nghiệm đa ô vuông la tinh

Một trong những hạn chế của thiết kế thí nghiệm theo kiểu LSD là có ít độ tự do và do đó ảnh hưởng đến độ chính xác của việc so sánh giữa các nghiệm thức. Ví dụ với 3 nghiệm thức thì chúng ta chỉ có 2 df ngẫu nhiên, trong khi đó với thiết kế thí nghiệm theo khối ngẫu nhiên hoàn toàn thì ta có 4 df với cùng số cá thể nghiên cứu và với 4 nghiệm thức ta chỉ có 6 df tự do so sánh với 9 df tự do trong trường hợp thiết kế thí nghiệm theo khối ngẫu nhiên hoàn toàn.

Để hạn chế khó khăn này trong khi vẫn giữ nguyên điểm mạnh không chế 2 yếu tố hàng và cột là sử dụng 2 ô vuông la tinh, có nghĩa là chúng ta gấp đôi số df ngẫu nhiên. Có hai cách để thực hiện được giải pháp này.

B	A	C
C	B	A
A	C	B

C	B	A
B	A	C
A	C	B

(a)

B	A	A	C	B	C
C	B	C	B	A	A
A	C	B	A	C	B

(b)

Biểu đồ 6.4. Hai hình thức thiết kế thí nghiệm đa ô vuông la tinh

(a) 2 hình vuông la tinh độc lập,

(b) hai hình vuông la tinh hỗn hợp để hình thành một hình chữ nhật 3x6

Trong biểu đồ 6.4a, hai hoặc nhiều hơn ô vuông được xem xét hoàn toàn độc lập với nhau, tính ngẫu nhiên hóa hoàn toàn độc lập lẫn nhau, giống như trong trường hợp ngẫu nhiên hóa của các nghiệm thức trong các khối khác nhau của một bố trí thí nghiệm theo khối ngẫu nhiên hoàn toàn.

Trong biểu đồ 6.4b, hai hay nhiều ô vuông la tinh có thể được trộn lẫn với nhau để hình thành một hình chữ nhật với mỗi nghiệm thức xuất hiện một lần trong cột và hai hoặc nhiều hơn trong mỗi hàng.

Sự khác nhau cần thiết giữa hai thiết kế thí nghiệm là trong (a) ảnh hưởng của một hàng cụ thể trong ô vuông thứ nhất là độc lập với ảnh hưởng của hàng tương ứng trong ô vuông thứ hai. Điều này cũng tương tự cho ảnh hưởng của cột. Trong hình chữ nhật, ảnh hưởng của hàng được giả sử là giống nhau cho cả hai ô vuông gốc và tương tự như vậy cho tất cả các cột. Ảnh hưởng của cột cho hai ô vuông gốc được giả sử là không có tương quan.

Các giả sử khác nhau dẫn đến sự khác nhau về hình thức phân tích. Đối với (a) tổng bình phương của các hàng và cột được tính một cách tách biệt cho các ô vuông. Một tổng bình phương của sự khác nhau tổng thể giữa hai ô vuông là được tính toán và tổng bình phương của các nghiệm thức được tính bằng cách dùng các tổng nghiệm thức của cả hai ô vuông.

Mô hình thống kê cho trường hợp a và b lần lượt thể hiện ở phương trình 6.25 và 6.26.

$$Y_{ijl} = \mu + K_l + R_{i(l)} + C_{j(l)} + T_k + e_{ijl} \quad [6.25]$$

với $l = 1, 2, \dots, s; i, j, k = 1, 2, \dots, t$

Trong đó:

- K_l là ảnh hưởng của ô vuông
- $R_{i(l)}$ là ảnh hưởng của hàng trong ô vuông
- $C_{j(l)}$ là ảnh hưởng của cột trong ô vuông
- T_k là ảnh hưởng của nghiệm thức
- e_{ijl} là ảnh hưởng của ngẫu nhiên

Bảng 6.25. ANOVA đối với một thí nghiệm LSD theo trường hợp ở hình 6.4a

Nguồn biến động	Tổng biến động	df
	$t^2 \sum (MG_l - GM)^2$	
Giữa các ô vuông	Hay $\frac{G_1^2 + G_2^2}{9} - \frac{\left(\sum x\right)^2}{18}$	$s - 1 = 1$
Hàng trong các ô vuông	$t \sum_i \sum_l (R_{i(l)} - MG_l)^2$ Hay $(SS_{\text{hàng}})_1 + (SS_{\text{hàng}})_2$	$s(t - 1) = 4$
Cột trong các ô vuông	$t \sum_j \sum_l (C_{j(l)} - MG_l)^2$ Hay $(SS_{\text{cột}})_1 + (SS_{\text{cột}})_2$	$s(t - 1) = 4$
Nghiem thức	$st \sum_k (T_k - GM)^2$ Hay $\frac{T_1^2 + T_2^2 + T_3^2}{6} - \frac{\left(\sum x\right)^2}{18}$	$t - 1 = 2$
Ngẫu nhiên	Bằng cách trừ $SS_{\text{tổng số}}$ cho các SS trên	$(st - s - 1)(t - 1) = 6$
Tổng số	$\sum_i \sum_j \sum_l (y_{ijl} - GM)^2$ Hay $\sum x^2 - \frac{\left(\sum x\right)^2}{18}$	$st^2 - 1 = 17$

Trong đó:

- MG_l là trung bình trong các ô vuông
- G_l là tổng trong các ô vuông
- $R_{i(l)}$ là trung bình của cột i trong ô vuông l
- $C_{j(l)}$ là trung bình của hàng j trong ô vuông l
- T_k là trung bình của nghiệm thức k
- GM là trung bình của tổng thể
- s là số hình vuông la tinh
- t là số nghiệm thức

$$Y_{ij} = \mu + R_i + C_j + T_k + e_{ij} \quad [6.26]$$

với $j = 1, 2, \dots, st; i, k = 1, 2, \dots, t$

Trong đó

- R_i là ảnh hưởng của hàng
- C_j là ảnh hưởng của cột
- T_k là ảnh hưởng của nghiệm thức
- e_{ij} là ảnh hưởng của ngẫu nhiên

Đối với (b), hình chữ nhật la tinh, tổng bình phương được tính là biến động giữa tổng các hàng, tổng các cột và tổng các nghiệm thức.

Bảng 6.26. ANOVA đối với một thí nghiệm LSD theo trường hợp ở hình 6.4b

Nguồn biến động	Tổng biến động	df
	$st \sum_i (R_i - GM)^2$	
Hàng	Hay $\frac{R_1^2 + R_2^2 + R_3^2}{6} - \frac{\left(\sum x\right)^2}{18}$	$t - 1 = 2$
	$t \sum_j (C_j - GM)^2$	
Cột	Hay $\frac{C_1^2 + C_2^2 + \dots + C_6^2}{3} - \frac{\left(\sum x\right)^2}{18}$	$st - 1 = 5$
	$st \sum_k (T_k - GM)^2$	
Nghiệm thức	Hay $\frac{T_1^2 + T_2^2 + T_3^2}{6} - \frac{\left(\sum x\right)^2}{18}$	$t - 1 = 2$
Ngẫu nhiên	Bằng cách trừ $SS_{\text{tổng số}}$ cho các SS trên	$(st - 2)(t - 1) = 8$
	$\sum_i \sum_j (y_{ij} - GM)^2$	
Tổng số	Hay $\sum x^2 - \frac{\left(\sum x\right)^2}{18}$	$st^2 - 1 = 17$

Sự khác biệt quan trọng nhất giữa hai hình thức phân tích ở trên là sự vượt trội về độ tự do trong trường hợp thiết kế thí nghiệm (b). Sự vượt trội này luôn luôn đúng cho mọi LSD có kích cỡ khác nhau.

Sai số của số trung bình và sai số của sự khác nhau giữa các trung bình được trình bày ở 6.27 và 6.28.

$$SEM = \sqrt{\frac{MS_E}{st}} \quad [6.27]$$

$$SED = \sqrt{\frac{2MS_E}{st}} \quad [6.28]$$

6.4. Bài tập

1. Một nhóm nghiên cứu muốn xem ảnh hưởng của các mức xơ thô (cao, trung bình, thấp) và mức protein (cao, trung bình, thấp) đến phát xạ NH_3 từ phân lợn. Nhóm nghiên cứu chia lợn ra làm 3 nhóm. Mỗi nhóm tương ứng với nghiệm thức như sau:
 - Nhóm 1: protein cao - xơ thô cao
 - Nhóm 2: protein trung bình - xơ thô trung bình
 - Nhóm 3: protein thấp - xơ thô thấp

Có trả lời được câu hỏi nghiên cứu về ảnh hưởng của mức xơ thô và mức protein đến phát xạ NH_3 từ phân lợn không?

2. Một nhóm nghiên cứu muốn đánh giá đáp ứng miễn dịch của một loại vắc xin A đối với dịch cúm gia cầm. Nhóm nghiên cứu quyết định áp dụng kiểu thiết kế thí nghiệm hình vuông la tinh. Hãy bình luận về quyết định sử dụng loại thiết kế thí nghiệm này để trả lời câu hỏi nghiên cứu?
3. Một nhóm nghiên cứu muốn đánh giá ảnh hưởng của 3 mức protein lên men (fermentable protein) trong khẩu phần đến sự phát xạ H_2S từ phân lợn. Nhóm nghiên cứu sử dụng 18 cá thể lợn giống nhau về các yếu tố tuổi, giống, giới tính, khối lượng... Mỗi nghiệm thức (mức protein lên men) được lặp lại 6 lần. Thí nghiệm được thiết kế theo kiểu CRD. Để phối hợp được khẩu phần có 3 mức protein lên men khác nhau nhóm đã sử dụng 3 nguồn protein khác nhau: bột cá, bột máu và khô dầu đậu tương. Mỗi nguồn protein cho mỗi mức protein lên men. Sau khi tiến hành thí nghiệm và xử lý số liệu, nhóm nghiên cứu kết luận mức protein lên men không ảnh hưởng đến sự phát xạ H_2S từ phân lợn. Hãy bình luận về kết luận của nhóm nghiên cứu
4. 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 kiểu hình 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 để thu thập số liệu. Hãy

- a. Xác định đơn vị thí nghiệm
 - b. Số lần lặp lại
 - c. Nhân tố thí nghiệm
 - d. Viết mô hình phân tích thống kê và giải thích các thành phần trong mô hình
5. Hãy bình luận nhận định “*Thiết kế thí nghiệm theo kiểu RCB làm tiêu tốn độ tự do*”?
 6. Hãy bình luận nhận định “*Khi thiết kế thí nghiệm chú ý mọi yếu tố phải tương đồng, ngoại trừ yếu tố thí nghiệm*”?
 7. Một nhóm quan tâm đến ảnh hưởng của 6 mức NSP, Polysaccharides phi tinh bột (NSP1, NSP2, NSP3, NSP4, NSP5, NSP6) khác nhau trong khẩu phần đến hàm lượng Cresols trong phân lợn (mg/kg). Cơ sở thí nghiệm có 24 con lợn cùng giống, cùng giới tính, cùng tuổi. Khối lượng ban đầu của lợn như sau:

68	42	33	51
42	37	68	72
51	33	54	52
42	57	35	56
50	46	62	48
54	36	42	44

Hãy đề xuất kiểu thiết kế thích hợp và lý giải vì sao?

8. Một nhóm nghiên cứu quan tâm đến ảnh hưởng của chế độ ăn mùa đông đến năng suất sữa của bò sữa. Có 4 loại khẩu phần (A, B, C và D) cần được kiểm tra ảnh hưởng của nó đến năng suất sữa của bò. Cơ sở thí nghiệm có 4 con bò.
Yếu tố thí nghiệm là gì? Có bao nhiêu mức của yếu tố thí nghiệm? Đơn vị thí nghiệm là gì? Số lần lặp lại là bao nhiêu? Nêu kiểu thiết kế thí nghiệm và nêu phương án phân tích số liệu?
9. Nhà nghiên cứu quan tâm đến ảnh hưởng của các chất kích thích sinh trưởng (ST1, ST2, ST3 và ST4) đến tăng trọng của lợn trong giai đoạn sinh trưởng. Có 4 loại chất kích thích. Có 24 cá thể lợn. Có 24 ô chuồng. Lợn đồng đều nhau về các yếu tố tuổi tác, trọng lượng, giới tính...
Yếu tố thí nghiệm là gì? Có bao nhiêu mức của yếu tố thí nghiệm? Đơn vị thí nghiệm là gì? số lần lặp lại là bao nhiêu? Nêu phương án thiết kế thí nghiệm và nêu phương án phân tích số liệu?
10. Một nhóm nghiên cứu quan tâm đến ảnh hưởng của 3 loại chế phẩm (A, B và C) đến hàm lượng của một chất X trong máu của lợn. Cơ sở thí nghiệm có 36 con lợn thuộc 3 giống khác nhau. Cơ sở thí nghiệm có đầy đủ cơ sở vật chất để tiến hành thí nghiệm. Hãy nêu kiểu thiết kế thí nghiệm và nêu phương án phân tích số liệu để kiểm tra ảnh hưởng của loại chế phẩm đến hàm lượng chất X trong máu.

11. Một thí nghiệm nghiên cứu ảnh hưởng của các mức bổ sung DL - Lysin khác nhau trong khẩu phần đến khả năng sản xuất của lợn thịt. Có 4 mức Lysin. Thí nghiệm được tiến hành trên 36 lợn thịt. Thiết kế thí nghiệm theo kiểu CRD. Cơ sở thí nghiệm có 9 ô chuồng. Mỗi ô chuồng nuôi 4 con lợn. Đơn vị thí nghiệm của thí nghiệm này là gì?
12. Một nhóm nghiên cứu muốn đánh giá ảnh hưởng của 3 mức protein khác nhau trong khẩu phần đến khả năng sản xuất của lợn. Thí nghiệm được tiến hành trên 18 con lợn thuộc 3 giống khác nhau. Các cá thể đồng đều về các yếu tố tuổi, khối lượng, giới tính... Nhóm nghiên cứu đã sử dụng kiểu thiết kế thí nghiệm CRD. Hãy bình luận về kiểu thiết kế thí nghiệm của nhóm nghiên cứu?
13. Trong trường hợp thí nghiệm một nhân tố, hãy nêu sự khác biệt về phương pháp ngẫu nhiên hóa các đơn vị thí nghiệm vào các nghiệm thức trong thiết kế thí nghiệm theo ngẫu nhiên hoàn toàn (CRD) và khối ngẫu nhiên đầy đủ (RCB)?

Chương 7

PHÂN TÍCH KHI BÌNH PHƯƠNG (χ^2)

Khi kiểm tra sự sai khác, phân tích phương sai áp dụng cho các nghiên cứu có biến phụ thuộc là biến định lượng, như tăng trọng (g/ngày), số con sơ sinh của lợn (con/ổ). Trong thực tế nghiên cứu (đặc biệt trong lĩnh vực thú y) chúng ta có thể có trường hợp giá trị của biến phụ thuộc là các số liệu phân nhóm. Ví dụ phương pháp vắt sữa (bằng tay hay bằng máy) ảnh hưởng đến tình trạng mắc bệnh viêm vú ở bò; biến phụ thuộc phân thành 2 nhóm: mắc bệnh và không mắc bệnh. Một ví dụ khác là tình hình mắc bệnh viêm vú (mắc bệnh và không mắc bệnh) ở bò sữa vào các mùa khác nhau (Mưa và Khô), biến phụ thuộc cũng phân thành 2 nhóm: mắc bệnh và không mắc bệnh. Hay bệnh viêm nội mạc tử cung ở bò có phụ thuộc vào giống bò hay không?

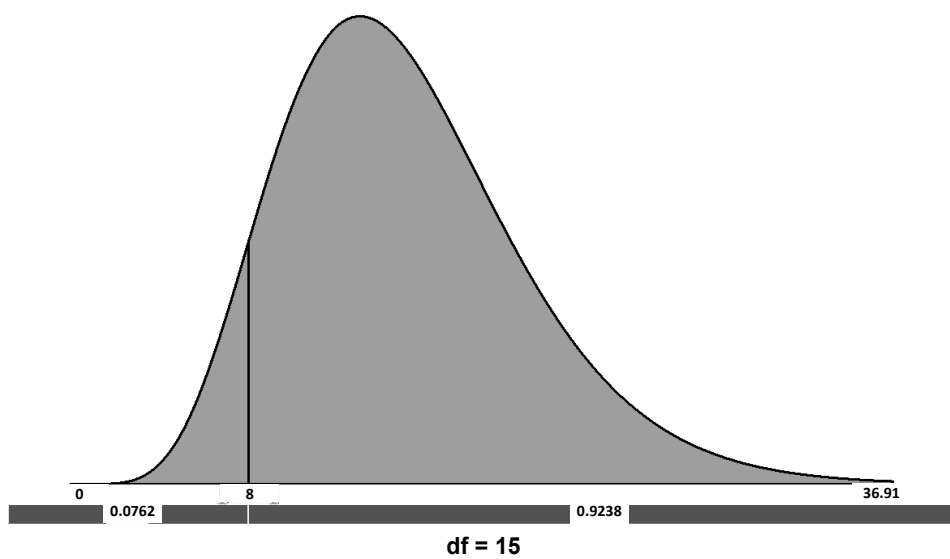
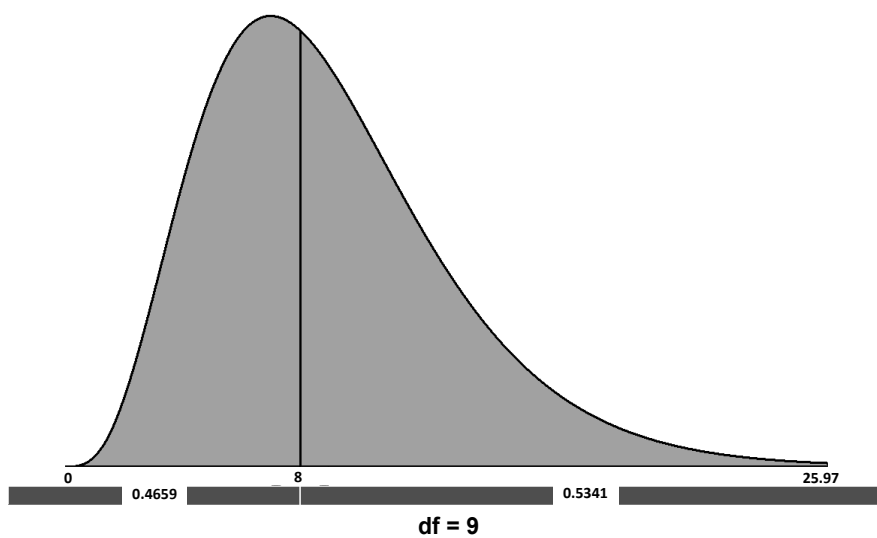
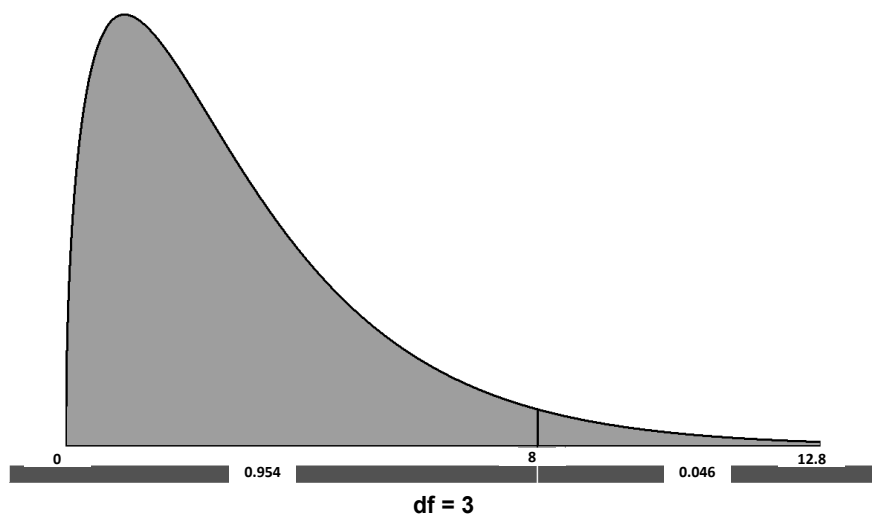
Trong thực tế chúng ta cũng có thể gặp tình huống kiểm tra sự phân bố lý thuyết (tuân theo các quy luật đã được chứng minh trước đây, ví dụ các quy luật di truyền Mendel) ví dụ kiểm tra tỷ lệ giới tính của quần thể có tuân theo tỷ lệ 1:1 hay không? hay kiểm tra sự phân ly của tính trạng theo các quy luật di truyền.

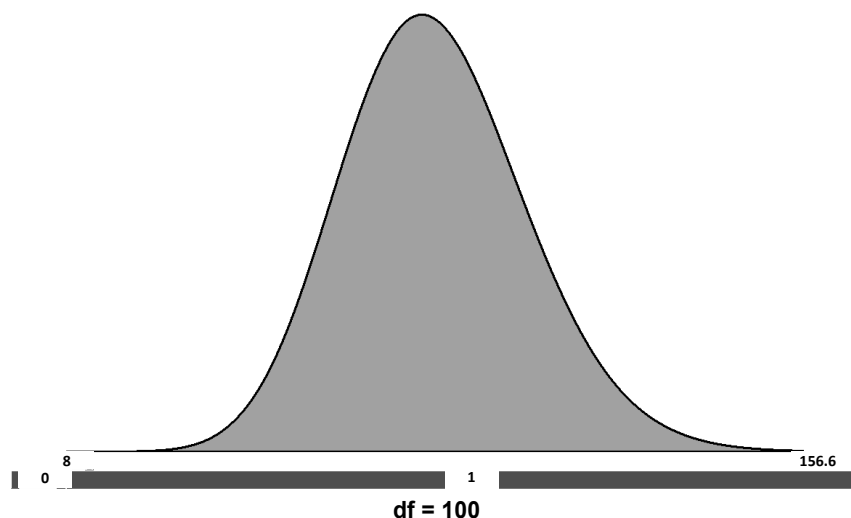
Phân tích khi bình phương có thể được sử dụng để giải quyết hai tình huống nêu trên. Chương này cung cấp các kiến thức liên quan đến phân bố khi bình phương, cách sử dụng phân tích khi bình phương để kiểm tra sự phân bố lý thuyết và kiểm tra sự phụ thuộc giữa các nhân tố hay tính trạng nghiên cứu.

7.1. Phân bố khi bình phương

Yếu tố phân biệt các phân bố khi bình phương chính là độ tự do (df). Mức độ tự do là một khái niệm rất quan trọng trong thống kê diễn dịch. Trong kiểm tra khi bình phương, độ tự do phụ thuộc vào số nhóm. Độ tự do rất quan trọng đối với phân bố khi bình phương, bởi vì mỗi độ tự do có một phân bố khi bình phương khác nhau. Chúng ta hãy xem xét biểu đồ 7.1 với 4 độ tự do khác nhau 3, 9, 15 và 100.

Khi độ tự do càng lớn thì sự phân bố càng ít lệch và toàn bộ sự phân bố chuyển sang bên phải của trục hoành. Phân bố khi bình phương khác với phân bố nhị thức ở chỗ nó là một phân bố liên tục hơn là một phân bố gián đoạn. Một biến trong một phân bố gián đoạn chỉ có thể nhận một số giá trị nhất định và có khoảng cách giữa các giá trị đó. Trong một phân bố xác suất gián đoạn, chúng ta có thể tìm được xác suất của mỗi phép thử/đầu ra.





Biểu đồ 7.1. Phân bố χ^2 với 4 mức độ tự do khác nhau

Ví dụ chúng ta có thể tính xác suất trong 5 lần gieo xúc xắc có 2 lần mặt sấp. Trong một phân bố liên tục, ví dụ như phân bố khi bình phương, chúng ta không thể tìm được giá trị xác suất không liên tục ví dụ $\chi^2 = 2$. Thay vào đó, chúng ta tìm được giá trị xác suất mà χ^2 giữa các giá trị cụ thể nào đó, ví dụ giữa 1,5 và 2,5 hay lớn hơn hay bé hơn một giá trị nào đó. Ví dụ $\chi^2 < 1,5$ hay $\chi^2 > 2,5$. Giả sử chúng ta quan tâm đến xác suất $\chi^2 > 8$. Chú ý rằng trong một đồ thị phân bố liên tục, "tỷ lệ của một vùng trong một phần nào đó là tương đương với xác suất của các giá trị trong vùng đó". Do vậy, $P(\chi^2 > 8)$ là tỷ lệ của vùng lớn hơn 8 trong phân bố khi bình phương. Chúng ta có thể thấy ở đồ thị 7.1. đường thẳng đứng tại $\chi^2 = 8$ cắt một phần rất nhỏ ở phía đuôi trên của phân bố khi bình phương với $df = 3$. Như vậy, với 3 độ tự do, $\chi^2 > 8$ có một xác suất nhỏ. Trong phân bố với 9 độ tự do cùng một đường cắt thẳng đứng như vậy, đã cắt một phần lớn hơn cho đuôi trên. Trong phân bố có độ tự do $df = 15$ thì phần lớn các phân bố nằm phía trên $\chi^2 = 8$. Và khi $df = 100$ gần như toàn bộ phân bố nằm phía trên đường $\chi^2 = 8$, như vậy ta thấy được sự cần thiết phải xem xét độ tự do trong việc đánh giá xác suất của một giá trị χ^2 .

Vùng cắt bởi các giá trị khác nhau của χ^2 với các độ tự do khác nhau được thể hiện trong bảng phân bố khi bình phương. Chúng ta sẽ thấy rằng các hàng của bảng là df , mỗi hàng cho chúng ta thông tin về phân bố khi bình phương với độ tự do df . Phần đầu của cột là tỷ lệ của vùng đuôi trên trong phân bố khi bình phương. Mỗi đầu vào trong bảng là một giá trị của χ^2 cắt phân bố tạo ra một tỷ lệ đuôi trên tương đương với phần đầu của cột. Giá trị χ^2 trong bảng được gọi là giá trị quyết định của χ^2 .

Bây giờ chúng ta hãy xem xét cách thức đọc thông tin từ bảng phân bố khi bình phương. Hãy xem xét giá trị quyết định/tới hạn trong bảng có $df = 6$ và cột là 0,05. Giá

trị quyết định là 12,59. Điều này có nghĩa là trong phân bố khi bình phương $df = 6$, một giá trị khi bình phương $\chi^2 = 12,59$ cắt 5% của vùng đuôi trên. Bởi vì, tỷ lệ của vùng trong một phần phân bố là xác suất của các giá trị trong phần phân bố đó, chúng ta có thể nói rằng trong phân bố khi bình phương với $df = 6$ xác suất $\chi^2 > 12,56 = 0,05$.

7.2. Kiểm tra sự phân bố lý thuyết

Kiểm tra sự phân bố lý thuyết hay còn gọi là kiểm định một phân phối.

Ví dụ 7.1. Giả sử bác sĩ A cho rằng 40% bệnh nhân thần kinh là do rối loạn trong suy nghĩ, 30% do rối loạn chức năng, 15% do rối loạn hành động, 10% do rối loạn hệ thần kinh và 5% thuộc các loại còn lại. Kết quả thực tế trong 200 bệnh nhân thần kinh tại một bệnh viện B cho rằng có 100 (50%) do rối loạn suy nghĩ, 60 (30%) do rối loạn chức năng, 20 (10%) do rối loạn hành động, 10 (5%) do rối loạn chức năng thần kinh và 10 (5%) thuộc các loại khác. Một câu hỏi đặt ra là "liệu kết quả từ bệnh viện B có trái ngược với nhận xét của bác sĩ A hay không?"

Để trả lời câu hỏi trên chúng ta tiến hành kiểm tra giả thuyết. Các bước của quá trình này hoàn toàn giống như các bước đã đề cập trong chương 4 (lực thống kê và kiểm tra giả thuyết). Với ví dụ 7.1 chúng ta có thể nói rằng: bác sĩ A là sai bởi vì ông cho rằng có 40, 30, 15, 10 và 5% các loại khác nhau của bệnh thần kinh, nhưng thực tế tại bệnh viện B những con số đó lại là 50, 30, 10, 5 và 5% những con số đó là không hoàn toàn giống nhau. Nhưng chú ý rằng kết quả của mẫu ít khi đồng nhất với giá trị của một quần thể (Ngoại trừ số phép thử tiến đến vô cùng). Nếu chúng ta chọn mẫu từ một quần thể với tỷ lệ con đực là 0,5, rất nhiều mẫu có tỷ lệ khác 0,5. Hay khi ta tung đồng xu 10 lần thì không phải lúc nào cũng có được 5 lần sấp và 5 lần ngửa. Thậm chí nếu bác sĩ A là đúng về nhận xét quần thể bệnh nhân thần kinh chúng ta vẫn có thể thấy được sự sai khác về phân bố của các loại bệnh nhân trong mẫu và trong quần thể.

Kiểm tra χ^2 được dùng để kiểm tra các giả thuyết của một quần thể có số liệu phân bố theo nhóm, căn cứ vào các số liệu của một mẫu từ quần thể. Điều này được thực hiện bằng cách so sánh tần suất quan sát trong mẫu và tần suất mong đợi nếu như giả thuyết là đúng. Nếu sự khác nhau đủ lớn thì chúng ta bác bỏ giả thuyết.

Tần suất mong đợi được ký hiệu là f_e (expected frequency); f_e trong một nhóm là tương đương với $N \times$ tỷ lệ mong đợi p_e (expected proportion).

$$f_e = Np_e \quad [7.1]$$

Kết quả f_e (theo nhận định của bác sĩ A) trong ví dụ 7.1 như sau:

Rối loạn suy nghĩ	$f_e = 200 \times 0,4 = 80$
Rối loạn chức năng	$f_e = 200 \times 0,3 = 60$
Rối loạn hành động	$f_e = 200 \times 0,15 = 30$
Rối loạn thần kinh	$f_e = 200 \times 0,1 = 20$
Nguyên nhân khác	$f_e = 200 \times 0,5 = 10$

Số lượng quan sát của mẫu thực tế thuộc vào các nhóm khác nhau thì được gọi là tần suất quan sát, ký hiệu là f_0 (Observed frequency); f_0 của 5 loại bệnh trên lần lượt là 100; 60; 20; 10 và 10.

Chúng ta có thể thấy rằng $f_0 \neq f_e$. Liệu nó khác nhau đủ để đưa ra kết luận phản bác giả thuyết? Chúng ta phải hình thành một cách đo lường sự khác nhau giữa f_0 và f_e .

Chúng ta bắt đầu bằng việc xem xét sự khác nhau giữa f_0 và f_e đó chính là $(f_0 - f_e)$ trong mỗi nhóm.

	Suy nghĩ	Chức năng	Hành động	Thần kinh	Khác	Σ
f_0	100	60	20	10	10	200
f_e	80	60	30	20	10	200
$f_0 - f_e$	20	0	- 10	- 10	0	0

Chúng ta chú ý rằng sự sai khác giữa tần suất quan sát và tần suất mong đợi là $= 0$, do vậy chúng ta không thể dùng $\sum (f_0 - f_e)$ như một đo lường sự khác nhau. Trong trường hợp này ta phải dùng giá trị bình phương, để cho tổng các chênh lệch giữa f_0 và f_e khác 0.

$(f_0 - f_e)^2$	400	0	100	100	0
-----------------	-----	---	-----	-----	---

Trong nhóm bệnh nhân rối loạn suy nghĩ, $(f_0 - f_e)^2$ là lớn hơn rất nhiều so với $(f_0 - f_e)^2$ trong nhóm rối loạn thần kinh. Nếu ta xét số lượng bệnh nhân thực tế, sự khác nhau giữa f_0 và f_e đối với rối loạn suy nghĩ là 100 so với 80 và đối với rối loạn hệ thần kinh là 10 đối với 20. Để hạn chế được hiện tượng này chúng ta chia $(f_0 - f_e)^2$ cho f_e .

5	0	3,33	5	0
---	---	------	---	---

$$\text{Lúc này ta có giá trị } \chi^2 = \sum \left[\frac{(f_0 - f_e)^2}{f_e} \right] \quad [7.2]$$

Như vậy, ta có thể thấy rằng f_0 và f_e càng khác xa nhau thì giá trị χ^2 càng lớn. Trong mẫu 200 bệnh nhân thần kinh tại bệnh viện B ta có giá trị

$$\chi^2 = 5 + 0 + 3,33 + 5 + 0 = 13,33$$

Bước tiếp theo là tìm giá trị xác suất của giá trị χ^2 đó là $P(\chi^2 > 13,33)$. Nếu xác suất là đủ nhỏ, chúng ta bác bỏ giả thuyết H_0 . Chúng ta đã tính được $\chi^2 = 13,33$. Giá trị $\chi^2 = 13,33$ còn gọi là giá trị tính toán. Để có giá trị tra bảng hay giá trị tới hạn chúng ta

cần xác định độ tự do (df). Trong kiểm tra sự phân bố lý thuyết, độ tự do ít hơn số nhóm một đơn vị hay chính là tổng số nhóm - 1. Ta tìm $P(\chi^2 > 13,33)$ với $df = 4$. Ta có P nằm trong khoảng 0,01 và 0,05 cho nên ta bác bỏ giả thuyết H_0 có nghĩa là giả thuyết của bác sĩ A là không phù hợp với số liệu thực tế ở bệnh viện B. Hay nói cách khác sự khác nhau giữa phân bố thực tế tại bệnh viện B và giả thuyết của bác sĩ A là có ý nghĩa tại mức 0,05.

Chúng ta có thể hiệu chỉnh giá trị χ^2 tính toán bằng cách sử dụng hệ số hiệu chỉnh Yates hay còn gọi là hiệu chỉnh tính liên tục cho bảng tương liên để loại bỏ sự thiên lệch. Giá trị χ^2 hiệu chỉnh như sau:

$$\chi^2 = \sum \frac{(|f_0 - f_e| - 0,5)^2}{f_e} \quad [7.3]$$

	Suy nghĩ	Chức năng	Hành động	Thần kinh	Khác
f_0	100	60	20	10	10
f_e	80	60	30	20	20
$f_0 - f_e$	20	0	- 10	- 10	0
$(f_0 - f_e - 0,5)^2$	380,25	0,25	90,25	90,25	0,25
$(f_0 - f_e - 0,5)^2 / f_e$	4,75	0,004	3,01	4,51	0,03

Giá trị χ^2 tính toán = 12,3 với $df = 4$, ta có P vẫn nằm trong khoảng 0,01 và 0,05 cho nên ta cũng bác bỏ giả thuyết H_0 .

7.3. Kiểm tra tính độc lập

Kiểm tra khi bình phương về tính độc lập (cũng có thể gọi là kiểm tra sự phụ thuộc) được dùng để kiểm tra liệu hai biến theo nhóm là liên quan với nhau hay phụ thuộc lẫn nhau hay độc lập với nhau hay không? Ví dụ giới tính (đực, cái) liên quan đến việc lựa chọn thức ăn khô hay ướt hay không? Tiêm vaccin hay không tiêm có liên quan đến tỷ lệ sống hay chết của gia súc hay không?

Thông thường số liệu của các nghiên cứu kiểm tra tính độc lập được tổ chức theo dạng bảng tương liên. Một biến ở hàng và biến còn lại ở cột (Bảng 7.1).

TR_i chính là tổng quan sát của sự kết hợp giữa các mức của Y với mỗi mức của X. TC_j là tổng quan sát của sự kết hợp giữa các mức của X với mỗi mức của Y. N chính là tổng các quan sát. f_0 là tần suất quan sát.

Bảng 7.1. Ví dụ về bảng tương liên

X = biến hàng	Y = biến cột				
	Y ₁	Y ₂	Y...	Y _l	TR _i
X ₁	f _{o11}	f _{o12}	...	f _{o1l}	TR ₁
X ₂	f _{o21}	f _{o22}	f _{o...}	f _{o2l}	TR ₂
...	...	...	...	...	...
X _k	f _{ok1}	f _{ok2}	f _{ok...}	f _{okl}	TR _k
TC _j	TC ₁	TC ₂	TC...	TC _l	N

Như đã đề cập ở trên, phân tích khi bình phương dựa vào sự sai khác giữa tần suất quan sát và tần suất mong đợi (hay còn gọi là tần suất lý thuyết bởi vì giá trị của nó được tính toán dựa vào giả thuyết H₀). Giả thuyết H₀ trong trường hợp kiểm tra tính độc lập là biến ở hàng không có quan hệ hay không phụ thuộc vào biến ở cột. Giả thuyết H_A là biến ở hàng và biến ở cột phụ thuộc lẫn nhau hay có quan hệ lẫn nhau hay không độc lập.

Giá trị của tần suất mong đợi (lý thuyết) được tính dựa vào giả thuyết H₀. Tần suất lý thuyết cho mỗi sự kết hợp biến giữa hàng và biến ở cột là:

$$f_{eij} = \frac{TR_i \times TC_j}{N} \quad [7.4]$$

Trong đó, f_{eij} là tần suất mong đợi trong mỗi sự kết hợp giữa biến hàng và biến cột. Giá trị χ^2 tính toán được tính theo công thức [7.5].

$$\chi^2 = \sum_1^k \sum_1^l \frac{(f_{oij} - f_{eij})^2}{f_{eij}} \quad [7.5]$$

Chúng ta so sánh giá trị χ^2 tính toán với giá trị χ^2 tra bảng. Độ tự do tra bảng = (số hàng - 1) (số cột - 1) hay (số mức của biến ở hàng - 1) × (số mức của biến ở cột - 1) = (k - 1)(l - 1). Nếu giá trị của χ^2 tính toán lớn hơn giá trị χ^2 tra bảng thì chúng ta bác bỏ giả thuyết H₀ ở mức α .

Ví dụ 7.2. Một nhóm nghiên cứu quan tâm đến liên quan giữa giới tính của lợn (đực, cái) và sự lựa chọn thức ăn (khô, ướt). Quan sát trên 71 con lợn (34 con cái và 37 con đực), kết quả lựa chọn thức ăn được thể hiện ở bảng 7.2. Hãy kiểm tra tính độc lập giữa giới tính với loại thức ăn.

Bảng 7.2. Bảng tương liên 2 × 2

		Thức ăn		Tổng hàng
		Khô	Ướt	
Giới tính	Cái	25	9	34
	Đực	17	20	37
Tổng cột		42	29	71 = N

Giả thuyết H_0 là không có mối quan hệ giữa giới tính và loại thức ăn. Nếu giả thuyết H_0 là đúng, chúng ta hy vọng rằng tỷ lệ con đực lựa chọn thức ăn khô và ướt tương tự như tỷ lệ con cái lựa chọn các loại thức ăn đó. Căn cứ vào công thức [7.4] ta có:

$$f_e(\text{cái, khô}) = \frac{34 \times 42}{71} = 20,1 \quad f_e(\text{đực, khô}) = \frac{37 \times 42}{71} = 21,9$$

$$f_e(\text{cái, ướt}) = \frac{34 \times 29}{71} = 13,9 \quad f_e(\text{đực, ướt}) = \frac{37 \times 29}{71} = 15,1$$

Để dễ dàng ước tính giá trị χ^2 tất cả f_e và f_0 được thể hiện trên cùng một bảng như bảng sau:

Bảng 7.3. Giá trị f_{0ij} và f_{eij}

		Khô	Ướt
Cái	f_0	25	9
	f_e	20,1	13,9
Đực	f_0	17	20
	f_e	21,9	15,1

Chú ý rằng tổng hàng và tổng cột của f_e là giống như tổng hàng và tổng cột của f_0 . Đây là tính tất yếu của kiểm tra tính độc lập của khi bình phương. Từ các giá trị ở bảng 7.3 và áp dụng công thức [7.5] giá trị χ^2 tính toán là = 5,614. Độ tự do để xác định giá trị χ^2 tới hạn là $(2 - 1) \times (2 - 1) = 1$. $P(\chi^2 > 5,61)$, $df = 1$, nằm giữa 0,05 và 0,01. Bởi vì $P < 0,05$ cho nên ta bác bỏ H_0 , sự khác nhau ở hai nhóm là ý nghĩa ở mức 0,05 hay chúng ta kết luận rằng con cái và con đực có khuynh hướng chọn lọc các loại thức ăn khác nhau, con cái có khuynh hướng thích thức ăn khô hơn.

Ví dụ 7.3. Để đánh giá hiệu quả sử dụng vacxin đối với một bệnh cụ thể trên lợn, một nhóm nghiên cứu đã tiến hành trên mẫu với 400 con lợn. Trước khi tiếp xúc với nguồn bệnh chọn 320 con được tiêm vacxin và 80 con làm đối chứng hay không được

tiêm vaccin. Sau đó cho tiếp xúc với nguồn bệnh và theo dõi số cá thể bị chết và sống, kết quả được thể hiện ở bảng 7.4. Nhóm nghiên cứu muốn biết liệu tiêm vaccin có ảnh hưởng đến tỷ lệ sống (chết) không?

Bảng 7.4. Bảng tương liên 2 x 2 trong thí nghiệm tiêm vaccin

		Sống/chết		Tổng hàng
		Sống	Chết	
Vaccin	Vaccin	199	121	320
	Đối chứng	60	20	80
Tổng cột		159	141	400 = N

Với phương pháp tiếp cận như đã nêu ở, các giá trị f_{0ij} và f_{eij} được thể hiện ở bảng 7.5.

Áp dụng công thức [7.5] ta có giá trị χ^2 tính toán = 4,603. Độ tự do để xác định giá trị χ^2 tới hạn là $(2 - 1) \times (2 - 1) = 1$. $P(\chi^2 > 4,603)$, $df = 1$, nằm giữa 0,05 và 0,01. Bởi vì $P < 0,05$ cho nên ta bác bỏ H_0 , sự khác nhau ở hai nhóm là ý nghĩa ở mức 0,05 hay chúng ta kết luận rằng tiêm vaccin cho lợn đã ảnh hưởng đến tỷ lệ sống, chết.

Bảng 7.5. Giá trị f_{0ij} và f_{eij}

	Tần suất	Chết	Sống
Đối chứng	f_0	20	60
	f_e	28,2	51,8
Vaccin	f_0	121	199
	f_e	112,8	207,2

Đối với trường hợp bảng tương liên có 4 ô có tần suất quan sát (f_0) lần lượt là a, b, c và d, giá trị χ^2 tính toán có thể được tính nhanh như sau:

a	b
c	d

$$\chi^2 = N \frac{(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \quad [7.6]$$

Áp dụng công thức [7.6] chúng ta cũng có giá trị χ^2 tính toán của ví dụ 7.3 là 4,603.

Chúng ta có thể hiệu chỉnh giá trị χ^2 tính toán bằng cách sử dụng hệ số hiệu chỉnh Yate hay còn gọi là hiệu chỉnh tính liên tục để loại bỏ sự thiên lệch. Giá trị χ^2 hiệu chỉnh như sau:

$$\chi^2 = \frac{\left(|ad - bc| - \frac{N}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)} N \quad [7.7]$$

Áp dụng công thức 7.7 ta có giá trị χ^2 tính toán hiệu chỉnh Yate là 4,06. Độ tự do để xác định giá trị χ^2 tới hạn là $(2 - 1) \times (2 - 1) = 1$. $P(\chi^2 > 4,06)$, $df = 1$, nằm giữa 0,05 và 0,01. Bởi vì $P < 0,05$ cho nên ta bác bỏ H_0 , sự khác nhau ở hai nhóm là ý nghĩa ở mức 0,05, hay chúng ta kết luận rằng tiêm vắc xin cho lợn đã ảnh hưởng đến tỷ lệ sống, chết.

Kiểm tra khi bình phương là không thích hợp khi $N \leq 20$, hay khi $20 < N \leq 40$ và có ô có tần suất lý thuyết < 5 . Khi xảy ra một trong hai trường hợp này, kiểm định chính xác Fisher được sử dụng. Khi $N \geq 100$ thì chúng ta sử dụng phép phân tích khi bình phương, khi $N < 100$ và không rơi vào hai trường hợp nêu trên thì nên sử dụng hiệu chỉnh Yate để giảm nhỏ giá trị χ^2 tính toán.

7.4. Kiểm tra chính xác Fisher đối với bảng tương liên 2×2

Kiểm tra chính xác Fisher hay còn gọi là kiểm tra Fisher - Freeman - Halton. Phép kiểm tra này dựa trên phân bố siêu hình học. Đặc trưng của phương pháp này là cho ra giá trị xác suất trực tiếp và căn cứ vào giá trị đó để quyết định chấp nhận hay bác bỏ giả thuyết H_0 . Kiểm tra chính xác Fisher tính trực tiếp giá trị xác suất, giá trị xác suất để có được một bảng tương liên lý thuyết như bảng tương liên quan sát.

Quay trở lại với bảng tương liên 2×2 ta có

a	b	a + b
c	d	c + d
a + c	b + d	N

Giá trị xác suất được tính theo công thức [7.8].

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} \quad [7.8]$$

Các bước để thực hiện kiểm tra chính xác Fisher tuân theo các bước sau:

- Từ tần suất quan sát tính giá trị p lần đầu tiên gọi là p_1
- Tính $ad - bc$

- Nếu $ad - bc > 0$ thì tăng a và d, đồng thời giảm b và c đúng một đơn vị sau đó lặp lại công thức [7.8] để tính giá trị p, còn gọi là p_2 (lý do là tính lần thứ 2); quá trình này lặp đi lặp lại cho đến khi a bằng giá trị tối thiểu (min) của $(a + b)$ hoặc $(a + c)$
- Nếu $ad - bc < 0$ thì giảm a và d, đồng thời tăng b và c đúng một đơn vị sau đó lặp lại công thức [7.8] để giá trị p, còn gọi là p_2 (lý do là lần tính thứ 2); quá trình này lặp đi lặp lại cho đến khi $a = 0$
- Tính giá trị xác suất P để từ đó quyết định chấp nhận hay bác bỏ H_0 .

$$P = 2(p_1 + p_2 + \dots P_n) \quad [7.9]$$

Nếu xác suất $P < 0,05$ thì bác bỏ giả thuyết H_0 (Nếu chúng ta lựa chọn $\alpha = 0,05$).

Ví dụ 7.4. Từ một đàn gia súc có 27 con, tiến hành tiêm vaccin cho 14 con và không tiêm vaccin cho 13 con, sau đó theo dõi, kết quả như sau:

Bảng 7.6. Bảng tương liên 2×2 trong thí nghiệm tiêm vaccin trong trường hợp sử dụng kiểm tra chính xác Fisher

		Sống/chết		Tổng hàng
		Sống	Chết	
Vaccin	Vaccin	12	2	14
	Đối chứng	3	10	13
Tổng cột		15	12	27 = N

Cùng ví dụ 7.4 nếu áp dụng kiểm tra χ^2 thì sẽ có giá trị χ^2 tính toán là 10,71 tương ứng với giá trị $P = 0,001$. Có nghĩa là chúng ta bác bỏ giả thuyết H_0 . Nếu chúng ta áp dụng hiệu chỉnh Yates (hiệu chỉnh về tính liên tục) thì giá trị χ^2 tính toán là 8,32, tương ứng với giá trị $P = 0,004$. Cũng có nghĩa là chúng ta bác bỏ giả thuyết H_0 . Nếu áp dụng kiểm tra chính xác Fisher chúng ta có giá trị $P = 0,002$. Mặc dù cả ba phương pháp đều cùng đưa đến kết luận bác bỏ giả thuyết H_0 nhưng giá trị P là khác nhau giữa ba phương pháp.

Cụ thể của phương pháp kiểm tra chính xác Fisher như sau:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} = \frac{14! \cdot 13! \cdot 15! \cdot 12!}{27! \cdot 12! \cdot 2! \cdot 3! \cdot 10!} = 0,00150$$

$ad - bc = 12 \times 10 - 3 \times 2 = 120 - 6 = 114 > 0$. Tiến hành tăng a và d đồng thời giảm b và c đúng một đơn vị.

		Sống/chết		Tổng hàng
		Sống	Chết	
Vaccin	Vaccin	12 + 1 = 13	2 - 1 = 1	14
	Đối chứng	3 - 1 = 2	10 + 1 = 11	13
Tổng cột		15	12	27 = N

Giá trị p ở lần tính tiếp theo là:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} = \frac{14! \cdot 13! \cdot 15! \cdot 12!}{27! \cdot 13! \cdot 2! \cdot 1! \cdot 1! \cdot 1!} = 0,00006$$

		Sống/chết		Tổng hàng
		Sống	Chết	
Vaccin	Vaccin	13 + 1 = 14	1 - 1 = 0	14
	Đối chứng	2 - 1 = 1	11 + 1 = 12	13
Tổng cột		15	12	27 = N

Giá trị p ở lần tính tiếp theo là:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} = \frac{14! \cdot 13! \cdot 15! \cdot 12!}{27! \cdot 14! \cdot 0! \cdot 1! \cdot 1! \cdot 12!} = 0,000001$$

7.5. Điều kiện sử dụng kiểm tra khi bình phương và kiểm tra chính xác Fisher

Các quan sát ở trong mỗi sự kết hợp giữa biến ở hàng và biến ở cột phải độc lập. Tính độc lập của các quan sát có ý nghĩa rằng mỗi quan sát chỉ rơi vào một ô (một sự kết hợp giữa biến ở hàng và biến ở cột) và chỉ một ô mà thôi. Trong ví dụ 7.2 nghiên cứu mối quan hệ giữa giới tính và việc lựa chọn thức ăn, một vài cá thể có thể ăn cả hai loại thức ăn (khô và ướt), như vậy n cá thể có thể được điền vào cả hai ô, nếu có điều này thì chúng ta không thể sử dụng kiểm tra khi bình phương hay kiểm tra chính xác Fisher.

Cũng như đã đề cập ở trên, kiểm tra χ^2 không nên dùng nếu f_e trong một ô hoặc một số ô là quá nhỏ. Một quy tắc là không nên dùng χ^2 khi một vài ô có $f_e < 5$ và có bất cứ ô nào có $f_e < 2$. Lý do đối với hạn chế này là xác suất phân bố χ^2 một phân bố liên tục. Chú ý rằng không có sự hạn chế đối với f_0 .

7.6. Bài tập

1. Một trung tâm thụ tinh nhân tạo tiến hành thử nghiệm 3 phương pháp thụ tinh nhân tạo khác nhau. Tỷ lệ phối có chữa ở 3 phương pháp thu được như sau: ở phương pháp I, 275 bò có chữa từ 353 bò tham gia thí nghiệm; tương tự ở phương pháp II, các con số này lần lượt là 192/256 con, ở phương pháp III là 261/384 con. Để kiểm tra có hay không sự khác nhau về tỷ lệ thụ tinh thành công ở 3 phương pháp này chúng ta dùng phép phân tích gì? Bản chất của phép phân tích là gì?
2. Một nhóm nghiên cứu về thời điểm (tháng trong năm) xảy ra dịch cúm H5N1 trên địa bàn tỉnh Thừa Thiên Huế. Nhóm nghiên cứu tiến hành điều tra thời điểm xảy ra dịch bệnh trên 167 thôn của các xã. Kết quả về thời điểm xảy ra dịch bệnh cũng như tần suất xuất hiện được thể hiện ở các thôn như sau.

Tháng	Số thôn xảy ra dịch bệnh
1	2
2	11
3	11
4	10
5	23
6	22
7	10
8	15
9	32
10	7
11	8
12	16
Tổng	167

Để kiểm tra xem thời điểm xảy ra dịch bệnh có khác nhau giữa các tháng hay không ta nên dùng phép kiểm tra gì? Bản chất của phép kiểm tra đó là gì? Hãy tiến hành phân tích sự phụ thuộc của việc xảy ra cúm gia cầm với thời điểm trong năm?

Chương 8

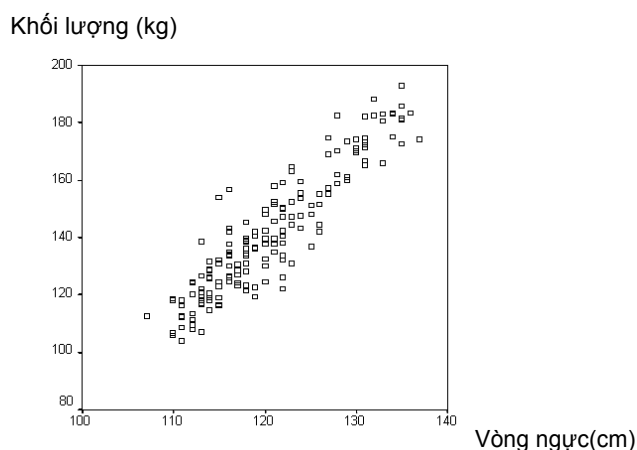
PHÂN TÍCH TƯƠNG QUAN HỒI QUY

Phân tích phương sai trả lời câu hỏi có hay không có ảnh hưởng của biến độc lập đến biến phụ thuộc, tuy nhiên không lượng hóa được sự ảnh hưởng đó. Các tính trạng thường có tương quan lẫn nhau, khoa học không chỉ dừng lại phát hiện có hay không có mối tương quan và khuynh hướng của mối tương quan mà còn yêu cầu lượng hóa các mối tương quan. Một tính trạng có thể bị ảnh hưởng bởi một hay nhiều nhân tố, các nhân tố đó có thể là định tính, định lượng hoặc cả hai. Khoa học cũng yêu cầu phân tích sự ảnh hưởng đó một cách định lượng. Phân tích tương quan hồi quy có thể đáp ứng các yêu cầu trên của nghiên cứu khoa học. Chương này trình bày lý thuyết về tương quan và phân tích tương quan hồi quy và chỉ đề cập đến hồi quy đơn biến.

8.1. Tương quan

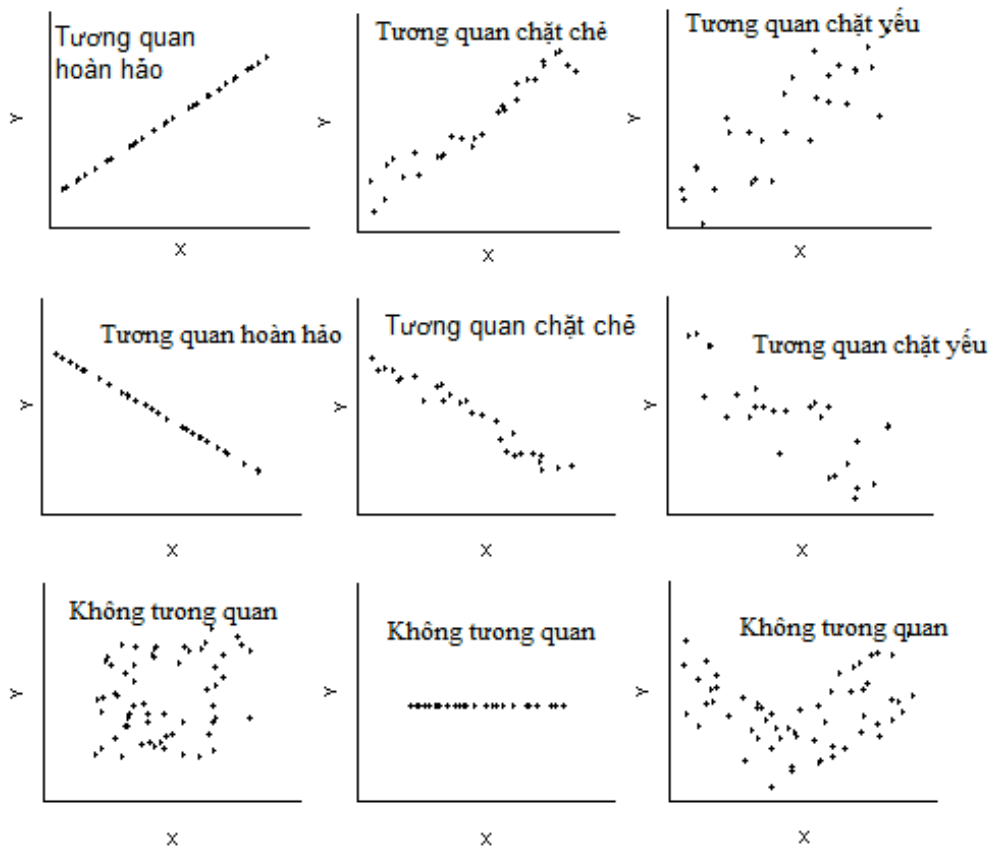
8.1.1. Biểu đồ phân tán và hiệp phương sai

Các tính trạng thường có mối tương quan lẫn nhau. Khi một tính trạng thay đổi thì tính trạng khác cũng có thể thay đổi. Sự thay đổi này có thể là dương tính hoặc âm tính (thuận hay nghịch). Sự thay đổi này còn được gọi là đồng biến (co - vary). Để có thể quan trắc mối tương quan giữa hai biến định lượng chúng ta có thể dùng Biểu đồ phân tán (scatter plot). Ví dụ 1 ở biểu đồ 8.1 biểu diễn mối quan hệ giữa vòng ngực (cm) và khối lượng của bò. Qua biểu đồ chúng ta có thể thấy rằng khi vòng ngực tăng thì khối lượng của bò cũng tăng.



Biểu đồ 8.1. Biểu đồ phân tán thể hiện mối tương quan giữa vòng ngực (cm) và khối lượng (kg) của bò

Mối tương quan giữa hai biến không chỉ theo khuynh hướng mà còn cả về sức mạnh. Sức mạnh của một mối tương quan là độc lập với khuynh hướng của nó; cả tương quan dương tính và âm tính đều có thể mạnh hoặc yếu. Về mặt hình thái, sức mạnh của một mối tương quan được thể hiện bằng các điểm trên biểu đồ phân tán tập trung xung quanh đường chéo từ phía thấp trái đến phía trên bên phải (tương quan dương) hoặc từ phía cao bên trái đến phía thấp bên phải (tương quan âm). Các điểm càng gần, càng sát đường chéo, mối tương quan giữa các biến càng chặt chẽ (biểu đồ 8.2).



Biểu đồ 8.2. Biểu đồ phân tán thể hiện sức mạnh mối tương quan giữa Y và X

Để lượng hóa sức mạnh của mối tương quan, chúng ta có thể dùng hiệp phương sai

$$\begin{aligned}
 \text{COV}_{xy} &= \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{N} \\
 &= \frac{(x_1 - \bar{X})(y_1 - \bar{Y}) + (x_2 - \bar{X})(y_2 - \bar{Y}) + \dots + (x_N - \bar{X})(y_N - \bar{Y})}{N} \\
 &= \frac{1}{N} \sum x_i y_i - \bar{x} \bar{y}
 \end{aligned} \quad [8.1]$$

Chú ý: N là số cặp các giá trị Y và X quan trắc trên cùng một đơn vị thí nghiệm.

Nếu xác định mức độ tương quan giữa hai tính trạng bằng hiệp phương sai thì giá trị của nó bị ảnh hưởng bởi đơn vị của tính trạng. Điều này có thể gây khó khăn khi so sánh sức mạnh tương quan giữa các loại tính trạng khác nhau.

8.1.2. Hệ số tương quan Pearson

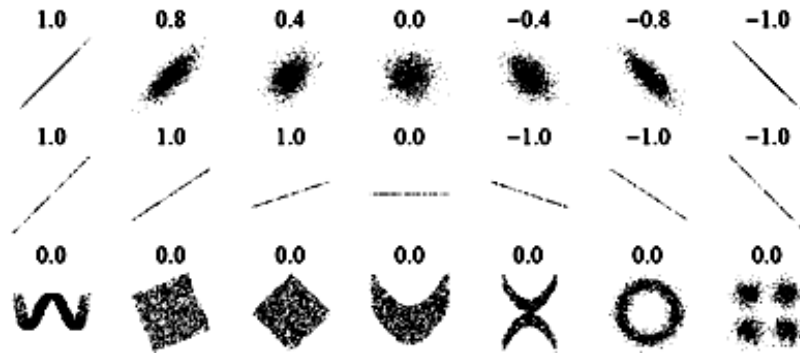
Sử dụng hiệp phương sai để đánh giá mức tương quan có nhược điểm. Nhược điểm này có thể được loại bỏ bằng cách làm mất đơn vị tính của hiệp phương sai. Phương pháp này có thể thực hiện bằng cách chia hiệp phương sai cho độ lệch chuẩn của cả hai biến. r chính là hệ số tương quan Pearson, nó là một chỉ số không có đơn vị, có thể so sánh mức độ tương quan giữa các nghiên cứu khác nhau hoặc giữa các cặp biến khác nhau. Về mặt hình thái học, hệ số tương quan Pearson có nghĩa là mức độ co cụm của số liệu trong việc hình thành nên đám mây phân bố tuyến tính.

$$r = \frac{\text{cov ar}_{xy}}{S_x \cdot S_y} = \frac{\text{cov ar}_{xy}}{\sqrt{(\text{var}_x)(\text{var}_y)}} \quad [8.2]$$

Từ công thức [8.2] ta có thể thấy hệ số tương quan là tỷ số giữa hiệp phương sai của hai biến và trung bình nhân của phương sai hai biến. Giá trị của hệ số tương quan nằm trong khoảng từ -1 đến 1. Nếu X và Y có tương quan dương thì r sẽ dương tính và ngược lại, r khi gần bằng 0 hoặc bằng 0 thì có nghĩa không có mối tương quan tuyến tính. Giá trị $|r|$ thể hiện sức mạnh của mối tương quan tuyến tính; càng gần đến ± 1 thì mối quan hệ càng mạnh, trong trường hợp tương quan hoàn hảo thì $r = \pm 1$. Dấu của hệ số tương quan chỉ ra hướng của mối tương quan. Khi r là dương thì đám đông phân bố cao dần từ trái sang phải. Khi r là âm thì đám đông phân bố thấp dần từ trái sang phải.

r	- 1	- 0,8	- 0,5	0	0,5	0,8	1
Tương quan	Âm, hoàn hảo	Âm, chặt chẽ	Âm, yếu	Không có tương quan tuyến tính	Dương, yếu	Dương, chặt chẽ	Dương, Hoàn hảo

Tránh nhầm lẫn giữa hướng của tương quan và sức mạnh của mối tương quan. Hướng và sức mạnh của mối tương quan độc lập lẫn nhau. Tương quan dương tính có thể là chặt chẽ hoặc yếu. Tương quan âm tính cũng có thể là chặt chẽ hoặc yếu.



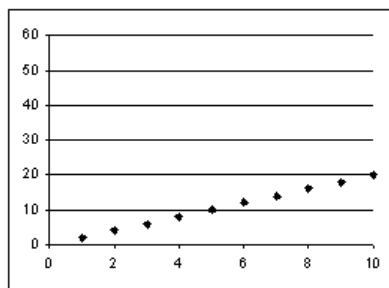
Biểu đồ 8.3. Biểu đồ biểu thị hướng và sức mạnh của hai biến x và y

Chúng ta có thể chứng minh giá trị của hệ số tương quan nằm trong khoảng từ -1 đến 1 bằng cách chuyển công thức [8.2] sang công thức của giá trị chuẩn hóa (z).

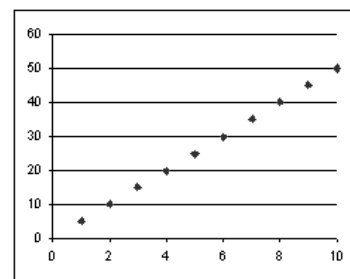
$$r = \frac{\text{covar}_{xy}}{S_x S_y} = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{N S_x S_y} = \frac{\sum \left(\frac{x_i - \bar{X}}{S_x} \right) \left(\frac{y_i - \bar{Y}}{S_y} \right)}{N} = \frac{\sum Z_x Z_y}{N} \quad [8.3]$$

Khi hai biến có một mối tương quan hoàn hảo, $Z_x = Z_y$ trong mỗi cặp. Do vậy, $Z_x Z_y$ trong mỗi cặp $= Z_x Z_x = Z_x^2 \Rightarrow r = \sum Z_x^2 / N$ chính là phương sai của giá trị Z, mà phương sai của giá trị Z là luôn luôn = 1. Do vậy, giá trị r của tương quan dương tính hoàn hảo là 1. Khi hai biến có mối tương quan âm tính hoàn hảo, $Z_x = -Z_y$ trong mỗi cặp $\Rightarrow Z_x Z_y = -Z_x^2 \Rightarrow r = - \left(\frac{\sum Z_x^2}{N} \right) \Rightarrow r = -1$.

Hệ số tương quan không đo lường độ dốc (hệ số hồi quy) của đường tuyến tính. Trong ví dụ của biểu đồ 8.3, cả hai trường hợp đều có hệ số tương quan hoàn hảo, $r = 1$; nhưng hệ số hồi quy là hoàn toàn khác nhau.



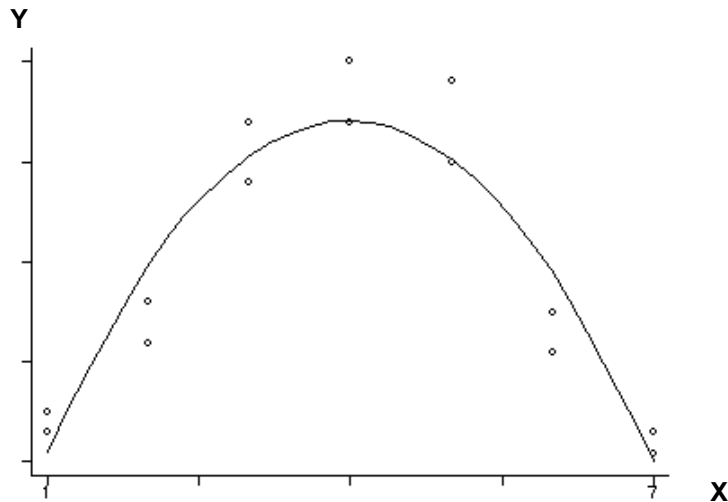
$r = 1,0$; hệ số hồi quy $(b) = 2$



$r = 1,0$; hệ số hồi quy $(b) = 5$

Biểu đồ 8.4. Hệ số hồi quy của hai phân bố khi có cùng $r = 1$

Hệ số tương quan chỉ đo lường sức mạnh của mối quan hệ tuyến tính giữa các biến. Tuy nhiên, cần chú ý rằng khi có một mối tương quan chặt chẽ giữa hai biến, điều này không luôn có nghĩa là hai biến này có mối quan hệ nguyên nhân kết quả. Muốn kiểm tra mối quan hệ nguyên nhân kết quả chúng ta phải thiết kế thí nghiệm và kiểm tra.

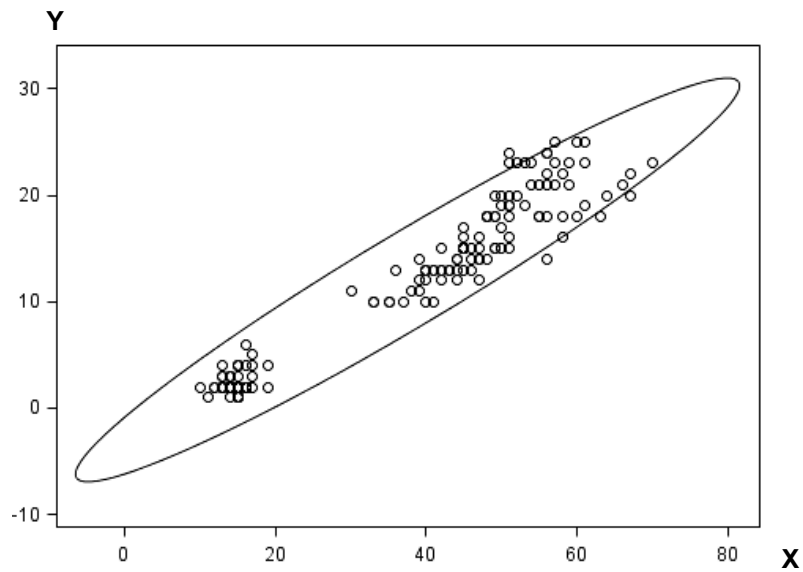


Biểu đồ 8.5. Mối quan hệ phi tuyến tính giữa Y và X

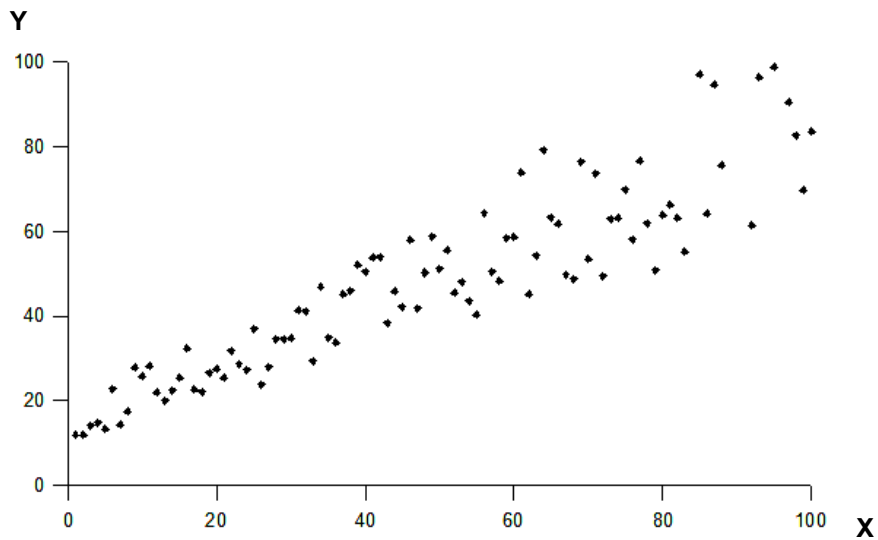
Khi các điểm (sự kết hợp giữa Y và X) trong một biểu đồ phân tán, tập trung thành một đường thẳng thì mối quan hệ đó được gọi là tuyến tính. Hệ số tương quan Pearson thể hiện mối quan hệ tuyến tính giữa hai biến. Nếu hai biến là không tuyến tính hơn là tuyến tính thì hệ số tương quan Pearson là ước tính không chính xác mối quan hệ giữa các biến. Ví dụ, biểu đồ 8.5 thể hiện mối quan hệ phi tuyến tính giữa Y và X. Nếu chúng ta tính hệ số tương quan Pearson thì $r = 0$. Tuy nhiên điều này không có nghĩa là giữa Y và X không có mối quan hệ mà giữa Y và X có thể có mối quan hệ chặt chẽ khác ví dụ parabol, nhưng không phải là quan hệ tuyến tính.

Hệ số tương quan Pearson đối với số liệu có mối tương quan phi tuyến tính không nhất thiết bằng 0 mà r là ước tính không chính xác sức mạnh của mối quan hệ, bởi vì r xác định mức độ mà các điểm phân tán thành một đường thẳng hơn là một đường cong. Do vậy trước khi xác định hệ số tương quan Pearson chúng ta nên xây dựng biểu đồ phân tán và quan trắc hình dạng của phân bố.

Trong phần lớn trường hợp tương quan tuyến tính, các điểm trong biểu đồ phân tán có khuynh hướng rơi vào trong hình elip. Mối quan hệ như vậy là được gọi là homoscedasticity (gốc từ Latin). Homo là giống nhau và scedasticity là phân tán. Nói cách khác homoscedasticity có nghĩa là độ phân tán (biến động) của các giá trị của Y là giống nhau ở các giá trị của X và độ phân tán của các giá trị của X là giống nhau với tất cả các giá trị của Y.



Biểu đồ 8.6. Biểu đồ phân tán dạng elip biểu thị quan hệ giữa Y và X



Biểu đồ 8.7. Biểu đồ phân tán thể hiện heteroscedasticity trong mối quan hệ giữa Y và X

Nhiều khi, các điểm của biểu đồ phân tán rơi vào trong một hình nón hoặc hình quả lê như biểu đồ 8.7. Biểu đồ này thể hiện một sự heteroscedasticity. Hetero là sai khác. Nói một cách khác, sự phân tán của giá trị Y là khác nhau tại các giá trị của X và sự phân tán của các giá trị của X là khác nhau tại các giá trị của Y.

Khi mà Heteroscedasticity tồn tại, sức mạnh của mối tương quan giữa các biến khác nhau thay đổi từ mức độ này sang mức độ kia. Ví dụ ở biểu đồ 8.7 mối tương quan giữa X và Y là mạnh khi X nhỏ nhưng lại yếu tại khi X cao. Trong trường hợp này, hệ số

tương quan Pearson xác định trung bình của mối quan hệ sức mạnh giữa X và Y cho tất cả các mức độ khác nhau giữa hai biến. Tuy nhiên nó là ước tính không chính xác sức mạnh của mối tương quan tại các mức độ khác nhau. Do vậy, r không phải là một chỉ số chính xác thể hiện mối quan hệ giữa các biến khi heteroscedasticity xuất hiện. Một lần nữa chúng ta thấy sự cần thiết phải xây dựng biểu đồ phân tán trước khi ước tính hệ số tương quan.

$r_{XY} = 0,50$ không có nghĩa là có 50% mối quan hệ giữa X và Y. r hoàn toàn không liên quan đến tỷ lệ. Tuy nhiên, r^2 có thể được diễn dịch như một tỉ lệ. Cụ thể r^2 là tỉ lệ của biến động (ví dụ phương sai) của mỗi biến được giải thích bởi biến khác. Ví dụ biến động của biến Y được giải thích bởi biến X. Khi $r = 0,5$ chúng ta có thể nói vài biến động của Y là cùng biến động với X và một vài biến động X là cùng biến động với Y. Do hệ số tương quan $\neq 1$ cho nên một phần biến động của mỗi biến là độc lập với biến khác. Bao nhiêu % sự biến động của mỗi biến là phụ thuộc lẫn nhau, lượng đó chính là $(0,5)^2 = 0,25$ hay 25% sự biến động của Y được giải thích bởi X, hay 25% sự biến động của X được giải thích bởi Y. Phần 75% còn lại của sự biến động của mỗi biến là độc lập với biến khác.

Bảng 8.1. Ảnh hưởng của hàm lượng protein trong khẩu phần đến phát xạ mùi từ phân lợn sinh trưởng

Protein ăn vào (g/kg thức ăn)	Nồng độ mùi (OE/m ³)
123	10211
142	28017
180	46430
123	5424
142	30985
180	32953
123	13246
142	3920
180	17739
123	3168
142	5030
180	22840
123	4867
142	53485
180	14982
123	13487
142	5871
180	110479

Nguồn: Lê Đình Phùng (2006)

Ví dụ 8.2. Kết quả nghiên cứu về ảnh hưởng của hàm lượng protein trong thức ăn đến phát xạ mùi từ phân lợn được trình bày ở bảng 8.1. Hệ số tương quan giữa hàm lượng protein trong thức ăn và phát xạ mùi từ phân lợn sinh trưởng được xác định là 0,52. Điều này có nghĩa là 27,2% sự biến động về phát xạ mùi từ phân lợn được giải thích bởi sự biến động về hàm lượng protein trong thức ăn. Còn lại 72,8% sự biến động chịu sự chi phối bởi các yếu tố khác.

Bởi vì r^2 là tương đương với tỉ lệ của phương sai của một trong hai biến bị ảnh hưởng bởi biến kia, r^2 được gọi là hệ số quyết định, tương tự như vậy $1 - r^2$, tỉ lệ của phương sai mà không bị ảnh hưởng bởi biến nghiên cứu, được gọi là hệ số không quyết định.

8.1.3. Kiểm tra giả thuyết về hệ số tương quan Pearson

Khi tính toán hệ số tương quan giữa hai biến X và Y chúng ta không chỉ mô tả mối tương quan của chúng ở mức độ mẫu mà chúng ta phải thể hiện được mối tương quan của chúng ở mức độ quần thể. Có nghĩa là chúng ta phải suy diễn, phải kiểm tra giả thuyết. Chúng ta sử dụng ρ là ký hiệu cho hệ số tương quan của quần thể và r là hệ số tương quan mẫu và ta dùng r để suy đoán ρ .

Về mặt giả thuyết, chúng ta có thể giả thuyết rằng ρ có giá trị nằm trong khoảng từ -1 đến 1. Tuy nhiên, trong thực tế thì kiểm tra về ρ lại kiểm tra giả thuyết H_0 mà X và Y là không có tương quan, có nghĩa là $\rho = 0$. Nếu bác bỏ H_0 thì ta có thể kết luận là hai biến có tương quan với nhau. Giống như các kiểm tra giả thuyết khác về trung bình của quần thể, kiểm tra $H_0: \rho = 0$ có thể kiểm tra một đuôi và hai đuôi. Nếu các giả thuyết nghiên cứu là X và Y là tương quan với nhau mà không nêu cụ thể hướng của tương quan thì kiểm tra hai đuôi được sử dụng.

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0$$

Nếu giả thuyết nghiên cứu là X và Y là tương quan dương thì kiểm tra thích hợp là kiểm tra một đuôi phía trên.

$$H_0: \rho \leq 0$$

$$H_A: \rho > 0$$

Nếu giả thuyết nghiên cứu là X và Y là có tương quan âm thì kiểm tra thích hợp là kiểm tra một đuôi phía dưới.

$$H_0: \rho \geq 0$$

$$H_A: \rho < 0$$

Các bước kiểm tra giả thuyết về ρ là giống như các giả thuyết kiểm tra khác:

1. Nêu H_0 và H_A .

2. Chọn mức ý nghĩa α .

3. Chọn mẫu từ quần thể, tốt hơn hết là chọn mẫu ngẫu nhiên, sau đó tính hệ số tương quan giữa 2 biến X và Y và tính t thống kê (giá trị t tính toán).

4. Giá trị tính toán của tham số thống kê là được so sánh với một hoặc nhiều giá trị quyết định để quyết định liệu xác suất của r mẫu giống như giá trị tính được là ít hơn hay lớn hơn α . Nếu $p < \alpha$ ta bác bỏ H_0 , tương quan giữa các biến là có ý nghĩa tại mức α và ta kết luận rằng các biến là có tương quan với nhau trong quần thể. Nếu $p > \alpha$ thì H_0 không bị bác bỏ có nghĩa tương quan là không có ý nghĩa tại mức α và chúng ta kết luận rằng các biến có thể không có tương quan trong quần thể.

Có hai phương pháp thay thế để tiến hành kiểm tra. Một phương pháp liên quan đến việc tính toán t từ số liệu mẫu và so sánh giá trị t tính toán với giá trị quyết định của t. Phương pháp khác bỏ qua việc tính toán t mà cho phép so sánh trực tiếp t với giá trị quyết định r.

- *Kiểm tra t*

Nếu X và Y là phân bố chuẩn trong quần thể và nếu H_0 là đúng thì tỉ lệ $\frac{r\sqrt{N-1}}{\sqrt{1-r^2}}$, tính toán cho tất cả các mẫu có thể, có kích thước N là phân bố t với $N - 2 = df$. Do vậy, kiểm tra thống kê trong kiểm tra t cho ý nghĩa của r là:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad [8.4]$$

Sau khi t được tính toán bằng công thức này, nó được so sánh với giá trị quyết định của t với $df = N - 2$.

Chúng ta hãy xem xét ví dụ sau:

Ví dụ 8.3. Theo một lý thuyết thì hai biến X và Y là có mối quan hệ âm tính. Giả sử X và Y là phân bố chuẩn. Một nhà nghiên cứu xác định 2 biến X và Y trên một mẫu 25 cá thể và có được $r = -0,32$, liệu kết quả này có ủng hộ lý thuyết không? Sử dụng $\alpha = 0,05$.

$$H_0: \rho \geq 0$$

$$H_A: \rho < 0$$

$$\text{Ta có } t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{-0,32\sqrt{25-2}}{\sqrt{1-(-0,32)^2}} = -1,62$$

Với $df = 23$ ta thấy rằng giá trị tuyệt đối của t nằm trong khoảng 1,319 và 1,714 với P lần lượt là 0,10 và 0,05 cho nên $P > \alpha$ có nghĩa là ta không bác bỏ giả thuyết H_0 . Giữa X và Y là không có mối tương quan âm ở mức $P < 0,05$.

- *Sử dụng bảng các giá trị quyết định của r*

Một cách đơn giản hơn để tiến hành kiểm tra mức ý nghĩa của r là sử dụng một bảng các giá trị quyết định của r.

Ví dụ 8.4. $r = 0,5$, $N = 20$ liệu mối tương quan này là sai khác có ý nghĩa so với giá trị 0 ở mức 0,05.

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0$$

$$\alpha = 0,05$$

Ta có $df = 18$, $r = 0,50$ nằm giữa 0,444 ($P = 0,05$) và 0,516 ($P = 0,01$) $\Rightarrow P < \alpha \Rightarrow$ bác bỏ $H_0 \Rightarrow$ hệ số tương quan mẫu là sai khác có ý nghĩa từ giá trị 0 ở mức 0,05 $\Rightarrow$ ta kết luận là hai biến có tương quan với nhau.

8.2. Phân tích hồi quy

8.2.1. Giới thiệu phương trình hồi quy

Hình thức tổng quát của phương trình hồi quy thể hiện mối quan hệ giữa biến X (độc lập) và biến Y (phụ thuộc) là:

$$\hat{Y} = a + bX \quad [8.5]$$

Phương trình hồi quy ở trên còn được gọi là phương trình hồi quy của Y lên X và được sử dụng cho việc ước tính Y từ X. Trong phương trình hồi quy, biến được ước tính (Y) còn được gọi là biến phụ thuộc hay biến đáp ứng và biến mà được sử dụng để ước tính biến được ước tính gọi là biến ước tính (X). Trong phương trình [8.5], b còn gọi là hệ số hồi quy của Y theo X và a là hằng số.

Ví dụ 8.5. Kết quả đánh giá khả năng sinh trưởng của bò lai Sind tại tỉnh Quảng Ngãi thu được phương trình sau:

$$\text{Khối lượng} = -209,81 + \text{Vòng ngực} \times 2,91 \quad [8.6]$$

Trong đó: Khối lượng (kg) được xác định bằng cân điện tử, vòng ngực được xác định bằng thước dây (cm). Từ phương trình [8.6] giá trị của khối lượng có thể được ước tính từ giá trị của vòng ngực.

Phương trình hồi quy [8.5] là dạng phương trình hồi quy đơn giản nhất. Tùy theo cách phân loại mà có thể có một số loại phương trình hồi quy như sau:

1. Phương trình hồi quy đơn biến: Phương trình hồi quy chỉ có một biến độc lập

Ví dụ 8.6. Khoảng cách lứa đẻ (ngày) = $406,8 + 0,631 \times$ tỷ lệ máu bò Brahman trong con mẹ (%) [8.7]

2. Phương trình hồi quy đa biến: Phương trình hồi quy có hơn một biến độc lập

+ Tất cả các biến độc lập đều là biến định lượng

Ví dụ 8.7. Khối lượng nạc = $7,231 + (\text{Khối lượng thân thịt nóng} \times 0,437 + (\text{Diện tích cơ thăn} \times 3,877) - (\text{Độ dày mỡ lưng giữa xương sườn 10 và 11} \times 18,746)$ [8.8]

Trong đó khối lượng đo bằng lb; diện tích đo bằng inch² (1 inch² = 6,5 cm²) và độ dày đo bằng inch.

Hay một ví dụ khác về ước tính pH của phân từ nồng độ ammonium và tổng số axit béo bay hơi trong phân

$$\text{pH} = 7,88 + 0,932 \text{ AM} - 0,50 \times \text{VFA} \quad [8.9]$$

Trong đó:

- AM là nồng độ ammonium trong phân (g/kg phân)
- VFA là nồng độ axit béo bay hơi tổng số trong phân (g/kg phân)
- pH là pH của phân

+ Một số biến độc lập là biến định lượng, trong khi đó một số biến là định tính

Ví dụ 8.8. Ảnh hưởng của một số nhân tố nghiên cứu đến tính trạng số con sơ sinh/lứa được thể hiện ở bảng dưới đây với các hệ số hồi quy bao gồm biến định tính và biến định lượng.

Biến	Mức của biến	Hệ số hồi quy
Giống lợn nái	Móng Cái (MC)	0,07
	F1 (MC × Y)	0
Đực giống	Landrace	0
	Yorkshire (Y)	- 0,38
Phương thức phối giống	Trực tiếp	0,24
	Nhân tạo	0
Phương thức nuôi	T.thống	- 0,95
	Bổ sung	0
Vùng sinh thái	Đồng bằng	0
	Trung du	- 0,70
	Ven biển	- 0,92
Chuồng trại	Đảm bảo	0
	Không ĐB	0,075
Mùa vụ	Hè	- 0,87
	Thu	- 0,002
	Xuân	0,14
	Đông	0
Lứa đẻ thứ		0,11

+ Trường hợp tất cả các biến độc lập là biến định tính thì còn gọi là phân tích phương sai.

Một dạng phân loại khác về phương trình hồi quy là:

- Phương trình hồi quy tuyến tính

Ví dụ 8.9. $I = 1,95 + 1,13 \times \ln(C)$ [8.10]

Trong đó: I là cường độ mùi phát xạ từ phân lợn;

C là nồng độ mùi từ phân lợn (OU_E/m^3).

- Phương trình hồi quy phi tuyến tính

Ví dụ 8.10. Phương trình ước tính khối lượng của bò F1 (Red Sindi $\times$ Vàng) từ vòng ngực có dạng:

$$Y_v = \frac{\alpha}{1 + \beta e^{-\gamma v}} \quad [8.11]$$

Trong đó Y là khối lượng (kg), v là vòng ngực (cm), α, β, γ là các tham số của mô hình.

8.2.2. Tiếp cận ước tính hệ số hồi quy

Ví dụ 8.11. Tập hợp số liệu ở bảng 8.2 được xác định trên 26 cá thể. Các giá trị từng cặp X và Y được xác định trên cùng một cá thể. Có nghĩa $N = 26$.

Biểu đồ phân tán của các cặp giá trị X và Y được thể hiện ở biểu đồ 8.8. Qua đó chúng ta thấy rằng mối quan hệ giữa X và Y có khuynh hướng tuyến tính. Áp dụng công thức [8.2], chúng ta xác định được hệ số tương quan Pearson giữa Y và X là 0,758. Áp dụng công thức [8.4], chúng ta có giá trị $t_{\text{tính toán}} = 5,69 > \text{giá trị } t_{0,025} = 2,069$. Điều này có nghĩa tồn tại mối quan hệ tuyến tính giữa Y và X. Với $r = 0,758$ chúng ta có thể kết luận khi tăng giá trị X sẽ làm tăng giá trị Y. Nhưng, chúng ta không thể trả lời được câu hỏi giá trị cụ thể của Y thay đổi một lượng bao nhiêu khi giá trị của X thay đổi một đơn vị.

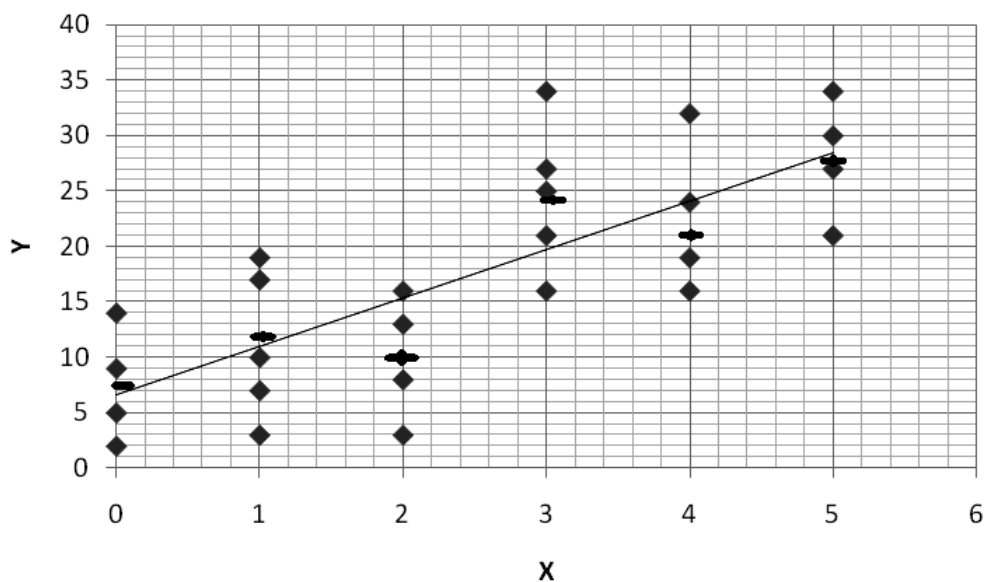
Bảng 8.2. Giá trị Y và X đo lường trên 26 cá thể khác nhau

Cá thể	X	Y	Cá thể	X	Y
1	3	34	14	0	14
2	1	7	15	2	16
3	4	19	16	1	3
4	2	3	17	5	21
5	1	19	18	4	16
6	0	5	19	5	34
7	5	30	20	3	21
8	4	32	21	3	16
9	1	10	22	4	24
10	2	8	23	0	9
11	3	25	24	5	27
12	0	2	25	3	27
13	1	17	26	2	13

Đối với mỗi giá trị của X, chúng ta ước tính giá trị của Y, nó là giá trị trung bình của Y cho tất cả các cá thể có giá trị X đó. Ví dụ khi X = 4, ta có giá trị của Y là

$$\frac{19 + 32 + 16 + 24}{4} = 22,75$$

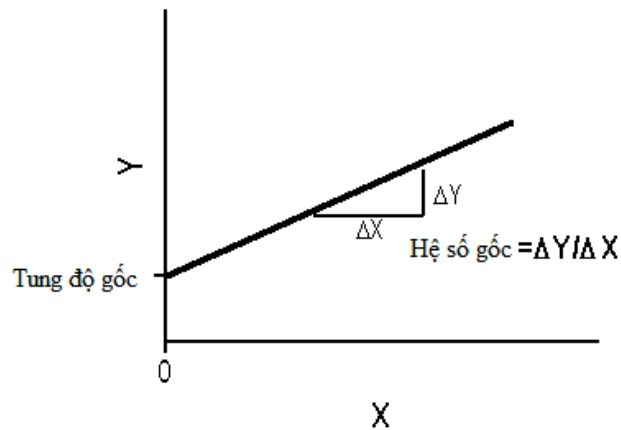
Tương tự như vậy chúng ta có giá trị của Y là 7,5; 11,2; 10; 24,6 và 28,25 tương ứng với các giá trị của X lần lượt là 0; 1; 2; 3 và 5. Trung bình của Y có khuynh hướng tăng khi X tăng. Tuy nhiên, giá trị thay đổi không phải là một hằng số từ giá trị X này sang giá trị X kia. Ví dụ: Sự khác nhau giữa trung bình Y đối với X = 1 và 2 là $-1,2 = 10 - 11,2$, nhưng sự khác nhau giữa trung bình Y đối với X = 2 và X = 3 là $14,60 = 24,6 - 10$. Do vậy, đường nối các trung bình của các dãy là bị dốc đặc. Khi N càng lớn thì đường nối giữa các giá trị trung bình của Y sẽ càng thẳng hơn, sẽ liền nét hơn (Biểu đồ 8.8). Đường liền đó là đường hồi quy dùng để ước tính Y từ mỗi giá trị của X. Nó là đường nối các giá trị trung bình Y khi các giá trị X thay đổi.



Biểu đồ 8.8. Biểu đồ phân tán thể hiện mối quan hệ giữa Y và X
(các nốt hình thoi là các giá trị quan trắc của Y ứng với mỗi giá trị của X và các nốt hình elip đen là trung bình của các giá trị Y ứng với mỗi giá trị của X)

Chúng ta có thể dùng đường hồi quy để ước tính giá trị của Y từ X. Để có thể làm được điều này, chúng ta xác định giá trị của X trên trục hoành, sau đó đối chiếu lên đường hồi quy và sau đó đối chiếu qua trục tung. Mặc dù ta có thể dùng phương pháp này nhưng chúng ta dùng phương pháp khác dễ dàng hơn. Chúng ta biết rằng bất cứ đường thẳng nào được vẽ từ cặp số liệu X và Y thì có một phương trình $Y = a + bX$, trong đó b là hệ số hồi quy và a là hằng số. Hệ số hồi quy của đường hồi quy đề cập đến lượng thay đổi của Y

đối với mỗi đơn vị thay đổi của X. Nói cách khác, đối với mỗi đơn vị tăng của X thì Y tăng lên bao nhiêu đơn vị. Hệ số này còn gọi là độ dốc của đường hồi quy. Nếu hệ số này là dương, Y tăng khi X tăng. Nếu hệ số này âm, Y giảm khi X giảm.



Biểu đồ 8.9. Hệ số góc và intercept trong phương trình hồi quy tuyến tính

8.2.3. Ước tính b và a của phương trình hồi quy

Từ biểu đồ 8.8, chúng ta thấy giá trị thực của Y (giá trị quan trắc) khác với giá trị Y ước tính ($\hat{Y}$) (giá trị nằm trên đường thẳng). Sự sai khác giữa giá trị quan trắc và giá trị ước tính của Y (ước tính dựa trên đường hồi quy) trên một cá thể/đơn vị thí nghiệm nào đó thì được gọi là sai sót ước tính hay hiệu dư.

$$Y_i - \hat{Y}_i = \text{Hiệu dư} \quad [8.12]$$

Trong nghiên cứu, hiệu dư là không tránh khỏi, ngoại trừ khi $r = \pm 1$. Tuy nhiên, đường hồi quy có thể được xác định theo một cách để hạn chế tối thiểu hiệu dư. Cụ thể là đường hồi quy được xây dựng theo một cách mà tổng của bình phương hiệu dư

$\sum (Y_i - \hat{Y}_i)^2$ là bé nhất. Vì lý do này, đường hồi quy còn được gọi là đường hồi quy bé nhất, hoặc còn gọi là đường phù hợp nhất. Mục tiêu của hồi quy tuyến tính là điều chỉnh giá trị của hệ số hồi quy và intercept để tìm đường hồi quy có thể ước tính tốt nhất các giá trị Y từ X.

Công thức tính a và b như sau:

$$b = r \left(\frac{S_y}{S_x} \right) \quad [8.13]$$

$$a = \bar{Y} - b\bar{X} \quad [8.14]$$

Từ ví dụ ở bảng 8.2, chúng ta có:

$$\Rightarrow b = r \left(\frac{S_y}{S_x} \right) = 0,758 \times \left(\frac{9,664}{1,669} \right) = 4,366$$

$$\Rightarrow a = \bar{Y} - b\bar{X} = 17,4^2 - 4,366 \times 2,46 = 6,64$$

Do vậy phương trình là: $\hat{Y} = 6,64 + 4,366X$

Chúng ta có thể tính a và b trực tiếp từ số liệu thô:

$$b = \frac{N(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{N(\sum X_i^2) - (\sum X_i)^2} \quad [8.15]$$

$$a = \frac{\sum Y_i - b(\sum X_i)}{N} \quad [8.16]$$

Phương trình hồi quy ước tính X từ Y là một phương trình khác so với phương trình ước tính Y từ X. Chúng ta không thể ước tính theo 2 hướng (Y từ X và X từ Y từ một phương trình hồi quy (ngoại trừ $r = \pm 1$). Phương trình hồi quy ước tính X từ Y có dạng:

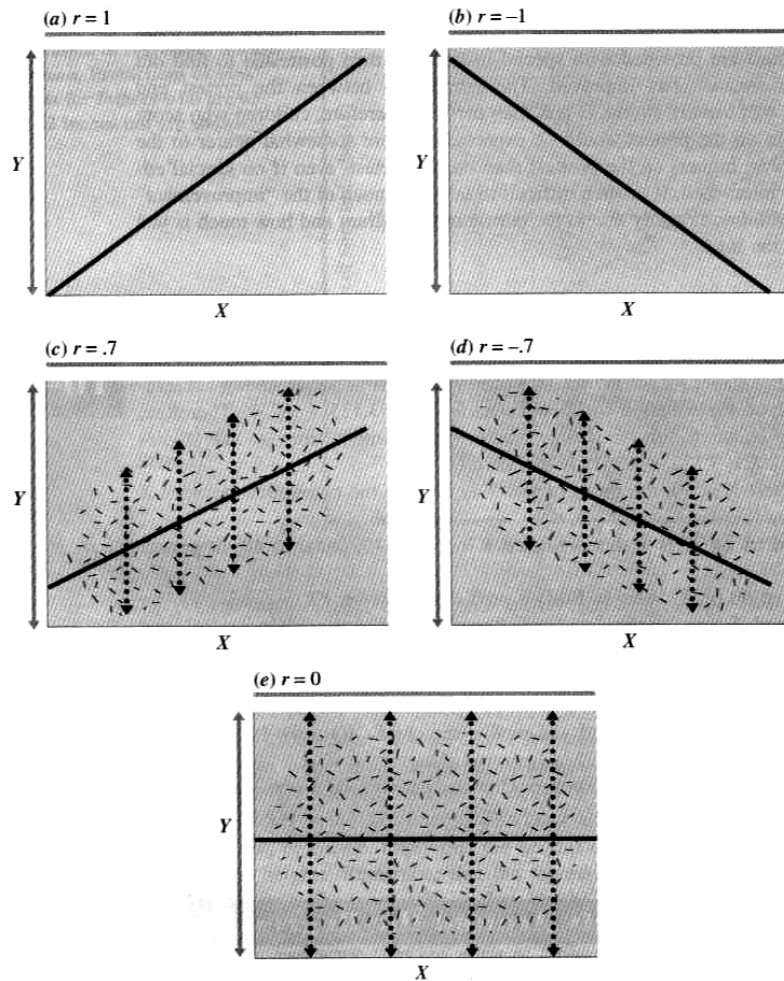
$$\hat{X} = a + bY \quad [8.17]$$

Chúng ta chỉ cần đảo vị trí của Y cho X và X cho Y trong công thức [8.15] và [8.16] thì sẽ có được công thức tính a và b cho phương trình hồi quy, công thức [8.17] và [8.18].

$$b = \frac{N(\sum X_i Y_i) - (\sum Y_i)(\sum X_i)}{N(\sum Y_i^2) - (\sum Y_i)^2} \quad [8.17]$$

$$a = \frac{\sum X_i - b(\sum Y_i)}{N} \quad [8.18]$$

Khi giá trị của biến ước tính là trung bình của X thì giá trị được ước tính là trung bình của Y. Nếu $r = 0$, giá trị ước tính của Y đối với tất cả giá trị của X là trung bình của Y. Khi $r = 0$ thì $b = 0$ có nghĩa đường hồi quy là một đường ngang, đường hồi quy cắt trục Y tại trung bình của Y. Và đường hồi quy lúc này là $\hat{Y} = \bar{Y}$. Như vậy khi $r = 0$, giá trị ước tính của Y cho tất cả giá trị của X là trung bình của Y.



Biểu đồ 8.10. Biến động của các giá trị quan trắc Y xung quanh đường hồi quy, mối liên quan với hệ số tương quan Pearson

Khi giá trị $|r|$ giảm thì biến động của giá trị quan trắc Y xung quanh giá trị ước tính tăng. Mối quan hệ giữa r và sự biến động của các giá trị quan trắc Y xung quanh giá trị ước tính được biểu diễn ở biểu đồ 8.10. Sự biến động của giá trị quan trắc Y xung quanh giá trị ước tính của Y được thể hiện bởi chiều dài của đường có 2 mũi tên. Khi $r = \pm 1$ tất cả các điểm trên biểu đồ phân tán nằm trên đường hồi quy và không có sự biến động của giá trị quan trắc của Y xung quanh giá trị ước tính của nó. Nói một cách khác, tất cả các giá trị ước tính là hoàn toàn đúng và tất cả hiệu dư = 0. Khi giá trị tuyệt đối của $r < 1$, các giá trị quan trắc của Y biến động xung quanh giá trị ước tính.

Phương sai của giá trị quan trắc của Y xung quanh giá trị ước tính của Y gọi là phương sai hiệu dư (S_E^2) và độ lệch chuẩn của các giá trị quan trắc Y xung quanh giá trị ước tính, gọi là sai số chuẩn hiệu dư (S_E). Phương sai hiệu dư là tương đương với tổng các bình phương hiệu dư chia cho N.

$$S_E^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{N} \quad [8.19]$$

$$S_E = \sqrt{S_E^2} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{N}} \quad [8.20]$$

Quay trở lại ví dụ ở bảng 8.2 ta có:

Tổng các bình phương hiệu dư: $\sum (Y_i - \hat{Y}_i)^2 = 1034,6786$

Phương sai hiệu dư: $S_E^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{N} = \frac{1034,6786}{26} = 39,79$

Sai số chuẩn của hiệu dư: $S_E = \sqrt{39,795} = 6,31$

Phương sai của hiệu dư = 39,79 nhỏ hơn so với phương sai của các giá trị quan trắc Y, $S_y^2 = 93,783$. Nói cách khác, các quan trắc Y ít biến động xung quanh giá trị Y ước tính hơn so với biến động xung quanh giá trị trung bình của tất cả các quan trắc. Điều này là hoàn toàn có thể giải thích bởi vì tương quan giữa X và Y là khá chặt chẽ (0,758). S_E^2 và S_E xác định sự biến động của các giá trị quan trắc Y xung quanh giá trị ước tính cho cả tập hợp số liệu. Nếu tất cả các dãy số liệu (tập hợp giá trị quan trắc Y ứng với mỗi giá trị X) có cùng mức độ biến động (homoscedasticity), phương sai của hiệu dư là phương sai của các giá trị quan trắc Y trong mỗi dãy số liệu và sai số chuẩn của hiệu dư như là độ lệch tiêu chuẩn của các giá trị Y trong mỗi dãy số liệu. Do vậy, nếu các giá trị quan trắc của Y là phân bố chuẩn thì 68% các giá trị quan trắc của Y nằm trong khoảng $1S_E$ của các giá trị ước tính của Y và 95% giá trị quan trắc của Y nằm trong khoảng $2S_E$ của giá trị ước tính của Y.

68% giá trị quan trắc Y nằm trong khoảng $\hat{Y} \pm S_E$

95% giá trị quan trắc Y nằm trong khoảng $\hat{Y} \pm 2S_E$

Xác định phương trình hồi quy không chỉ dừng lại ở mẫu mà thường phải khái quát hóa ở mức quần thể. Hồi quy mẫu cho việc ước tính Y từ X cần 2 tham số: hệ số hồi quy b và intercept a. Giá trị của b và a là biến động từ mẫu này sang mẫu nọ. Điều này cũng có nghĩa, đường hồi quy là khác nhau từ mẫu này sang mẫu nọ. Tuy nhiên, b và a là ước tính không chệch của hệ số hồi quy và intercept. Do vậy, phương trình hồi quy mẫu, $\hat{Y} = a + bX$ là ước tính tốt nhất của phương trình hồi quy của quần thể. Tuy nhiên phương sai hiệu dư của quần thể cần được hiệu chỉnh, điều này có nghĩa sai số chuẩn của hiệu dư cũng cần hiệu chỉnh.

$$s_E^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{N - 2} \quad [8.21]$$

$$\text{và } s_E = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{N - 2}} \quad [8.22]$$

Phương sai hiệu dư ước tính phụ thuộc vào $\hat{Y}$, mà $\hat{Y}$ lại phụ thuộc vào b và a do vậy độ tự do là N - 2. Tuy nhiên khi N lớn thì việc ước tính với độ tự do bằng N hay N - 2 không sai khác nhau nhiều.

8.2.4. Khoảng tin cậy của giá trị ước tính của Y

Nếu các giá trị của Y được giả thuyết là phân bố chuẩn và nếu quan hệ giữa X và Y là đồng nhất về phương sai, chúng ta có thể ước tính khoảng tin cậy của các giá trị ước tính của Y hay chính là khoảng biến động của giá trị ước tính của Y với mức độ tin cậy α nhất định.

$$PI = \hat{Y} \pm t_{\alpha/2} s_E \sqrt{1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{SS_X}} \quad [8.23]$$

Trong đó:

- PI: khoảng tin cậy của giá trị ước tính
- $\hat{Y}$: giá trị ước tính của Y tại X_i
- $t_{\alpha/2}$: giá trị quyết định của t trong kiểm tra hai đuôi với mức α và $df = N - 2$
- s_E : sai số chuẩn của hiệu dư
- SS: tổng các bình phương các độ lệch của X

Ví dụ: xem xét số liệu ở bảng 8.2. Giả sử chúng ta muốn ước tính khoảng biến động 95% của giá trị ước tính của Y

Ta có $df = 24$; $\alpha = 0,05$ có nghĩa $t_{\alpha/2} = 2,064$ (tra bảng t). $s_E = 6,57$ (áp dụng công thức 8.22)

$$\bar{X} = 2,46 \quad SS_X = 72,46$$

$$PI = \hat{Y} \pm 2,064 \times 6,57 \sqrt{1 + \frac{1}{26} + \frac{(X_i - 2,46)^2}{72,46}}$$

Các giá trị $\hat{Y}$ và X_i thay đổi từ đơn vị thí nghiệm này sang cá thể khác. X_i là X của cá thể i và $\hat{Y}$ là giá trị ước tính từ phương trình hồi quy đã được xác định $\hat{Y} = 6,59 + 4,40X$.

Chúng ta hãy tính khoảng ước tính 95% cho 2 cá thể A và B. Cá thể A có $X = 5$, cá thể B có $X = 3$. Trước hết ta sử dụng $\hat{Y} = 6,59 + 4,4X$ để ước tính $\hat{Y}$ cho hai cá thể A và B.

$$\text{Cá thể A: } \hat{Y} = 6,59 + 4,40(5) = 28,59$$

$$\text{Cá thể B: } \hat{Y} = 6,59 + 4,40(3) = 19,79$$

Có nghĩa là khoảng ước tính 95% cho cá thể A là:

$$PI = 28,59 \pm 13,56 \sqrt{1 + \frac{1}{26} + \frac{(5 - 2,46)^2}{72,46}} = 28,59 \pm 14,399$$

Khoảng ước tính 95% cho cá thể B là:

$$PI = 19,79 \pm 13,56 \sqrt{1 + \frac{1}{26} + \frac{(3 - 2,46)^2}{72,46}} = 19,79 \pm 13,854$$

Như vậy, chúng ta 95% tin tưởng rằng cá thể A có giá trị ước tính $\hat{Y}$ nằm trong khoảng 14,19 và 42,99 và cá thể B có giá trị $\hat{Y}$ ước tính nằm trong khoảng 5,945 và 33,635.

8.2.5. Các thành phần phương sai và hệ số quyết định

Trong chương 4, chúng ta đã thảo luận việc tách tổng biến động thành những biến động thành phần. Chúng ta có thể sử dụng nguyên lý tương tự để xác định hệ số quyết định.

Chúng ta có:

$$SS_T = \sum (Y_i - \bar{Y})^2 \quad [8.24]$$

SS_T là tổng các bình phương các độ lệch tổng số, gọi tắt là tổng các bình phương tổng số.

$$SS_{\hat{Y}} = \sum (\hat{Y} - \bar{Y})^2 \quad [8.25]$$

$SS_{\hat{Y}}$ là tổng các bình phương do phương trình hồi quy. $\hat{Y} - \bar{Y}$ là một lượng biến động trong các giá trị quan trắc Y được giải thích bởi X , còn gọi là phần biến động được giải thích.

$$SS_E = \sum (Y_i - \hat{Y})^2 \quad [8.26]$$

SS_E là tổng các bình phương hiệu dư, còn gọi là phần biến động không được giải thích.

Chúng ta có:

$$SS_T = SS_{\hat{Y}} + SS_E \quad [8.27]$$

Hệ số quyết định có dạng tổng quát là:

$$R^2 = \frac{SS_{\hat{Y}}}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad [8.28]$$

Quay trở lại ví dụ ở bảng 8.2, ta có

$$R^2 = 1 - \frac{1034,68}{2438,35} = 0,5757$$

Điều này có nghĩa khoảng 58% biến động của Y là được giải thích bởi X, phần còn lại 42% được giải thích bởi các biến khác mà không biết được. Nếu ta tính căn bậc 2 của 0,5753 ta có 0,76, đây chính là hệ số tương quan.

Bảng 8.3. Các thành phần tổng các bình phương từ tập hợp số liệu ở bảng 8.2

	X_i	Y_i	$(Y_i - \bar{Y})^2$	$\hat{Y}_i$	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \hat{Y}_i)^2$
	3	34	274,79	19,79	5,60	201,92
	1	7	108,64	10,99	41,38	15,92
	4	19	2,49	24,19	45,79	26,94
	2	3	208,03	15,39	4,13	153,51
	1	19	2,49	10,99	41,38	64,16
	0	5	154,33	6,59	117,36	2,53
	5	30	158,18	28,59	124,70	1,99
	4	32	212,49	24,19	45,79	61,00
	1	10	55,10	10,99	41,38	0,98
	2	8	88,79	15,39	4,13	54,61
	3	25	57,41	19,79	5,60	27,14
	0	2	237,87	6,59	117,36	21,07
	1	17	0,18	10,99	41,38	36,12
	0	14	11,72	6,59	117,36	54,91
	2	16	2,03	15,39	4,13	0,37
	1	3	208,03	10,99	41,38	63,84
	5	21	12,79	28,59	124,70	57,61
	4	16	2,03	24,19	45,79	67,08
	5	35	308,95	28,59	124,70	41,09
	3	21	12,79	19,79	5,60	1,46
	3	16	2,03	19,79	5,60	14,36
	4	24	43,26	24,19	45,79	0,04
	0	9	70,95	6,59	117,36	5,81
	5	27	91,72	28,59	124,70	2,53
	3	27	91,72	19,79	5,60	51,98
	2	13	19,56	15,39	4,13	5,71
Trung bình	2,46	17,42				
Tổng			2438,35		1402,86	1034,68

Hệ số R^2 đo lường mức độ sát thực của đường hồi quy với các số liệu quan trắc. Khi $R^2 = 1$ đường hồi quy hoàn toàn phù hợp với số liệu quan trắc hay mô hình hồi quy có thể giải thích được toàn bộ sự biến động của Y . Giá trị của R^2 có thể nằm ngoài khoảng $[0, 1]$. Khi $R^2 = 0$ có nghĩa là không có mối quan hệ tuyến tính.

Hệ số R^2 điều chỉnh (còn có ký hiệu $\bar{R}^2$) là một dạng biến thể của R^2 trong đó có điều chỉnh số lượng biến giải thích (X) trong mô hình. Không giống như R^2 , $\bar{R}^2$ điều chỉnh chỉ tăng nếu yếu tố giải thích mới thêm vào thực sự làm tăng khả năng giải thích của mô hình. $\bar{R}^2$ điều chỉnh có thể là âm tính và sẽ luôn luôn nhỏ hơn hoặc bằng R^2 . Công thức tính $\bar{R}^2$ điều chỉnh được thể hiện ở [8.29].

$$1 - (1 - R^2) \frac{N - 1}{N - p - 1} = 1 - \frac{SS_E}{SS_T} \frac{df_T}{df_E} \quad [8.29]$$

Trong đó p là tổng số biến giải thích trong mô hình hồi quy tuyến tính; N là kích thước mẫu. Thực tế $\bar{R}^2$ điều chỉnh là:

$$\bar{R}^2 = 1 - \frac{S_E^2}{S_T^2} \quad [8.30]$$

8.2.6. Kiểm tra giả thuyết phương trình hồi quy và các hệ số của phương trình hồi quy

Khi xây dựng phương trình hồi quy nói chung và phương trình hồi quy đơn biến nói riêng, chúng ta cần kiểm tra sự tồn tại của phương trình hồi quy. Sự tồn tại của phương trình hồi quy tuyến tính đơn biến phụ thuộc vào hệ số a và b .

- Giả thuyết H_0 để kiểm tra sự tồn tại của phương trình hồi quy tuyến tính đơn biến là $a = 0$ và $b = 0$
- Giả thuyết H_A ít nhất một trong hai tham số a hoặc b khác 0

Kiểm tra giả thuyết H_0 dựa vào kiểm tra F với giá trị $F_{\text{tính toán}}$

$$F_{\text{tính toán}} = \frac{MS_{\text{hồi quy}}}{MS_E} \quad [8.31]$$

$$MS_{\text{hồi quy}} = \frac{SS_{\text{hồi quy}}}{df_{\text{hồi quy}}} = \frac{SS_{\hat{Y}}}{df_{\text{hồi quy}}} \quad [8.32]$$

$$MS_E = \frac{SS_E}{df_E} \quad [8.33]$$

Giá trị F tính toán được so sánh với giá trị của F quyết định hay F tra bảng với độ tự do từ số là $df_{\text{hồi quy}}$ và độ tự do mẫu số là df_E ở mức ý nghĩa α . Nếu giá trị $F_{\text{tính toán}} > =$

$F_{\text{quyết định}}$ chúng ta bác bỏ giả thuyết H_0 có nghĩa là phương trình hồi quy tồn tại và ngược lại.

Quay trở lại ví dụ 8.11 chúng ta có kết quả kiểm tra sự tồn tại của phương trình hồi quy ở bảng 8.4.

Bảng 8.4. Kết quả kiểm tra sự tồn tại của phương trình hồi quy ở ví dụ 8.11

Nguồn biến động	SS	df	MS	F	P
Hồi quy	1403,67	1	1403,67	32,56	0,000
Hiệu dư	1034,69	24	43,11		
Tổng số	2438,35	25			

Trong phương trình hồi quy tuyến tính đơn biến thì độ tự do của phương trình hồi quy luôn luôn bằng 1. Kết quả bảng 8.4 cho thấy giả thuyết H_0 bị bác bỏ. Có nghĩa phương trình hồi quy thực sự tồn tại.

Khi phương trình hồi quy tồn tại, chúng ta cần kiểm tra sự tồn tại của hệ số hồi quy (b).

- Giả thuyết H_0 : $b = 0$
- Giả thuyết H_A : b khác 0

Chúng ta sử dụng kiểm tra t để kiểm tra giả thuyết H_0 : $b = 0$. Giá trị $t_{\text{tính toán}}$ được ước tính theo công thức [8.34].

$$t_b = \frac{b}{SE_b} \quad [8.34]$$

Giá trị của b được ước tính bởi công thức [8.17] SE_b được ước tính theo công thức [8.35]

$$SE_b = \frac{SE}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2}} \quad [8.35]$$

SE được ước tính theo công thức [8.20].

Giá trị t_b được so sánh với giá trị $t_{\text{tra bảng}}$ với độ tự do df_E và mức ý nghĩa $\alpha/2$ (kiểm tra hai đuôi). Nếu $t_b \geq t_{\text{tra bảng}}$, chúng ta bác bỏ H_0 có nghĩa hệ số b tồn tại (khác 0).

Chúng ta cũng có thể kiểm tra sự tồn tại của hệ số a (tung độ gốc)

- Giả thuyết H_0 : $a = 0$
- Giả thuyết H_A : a khác 0

Chúng ta sử dụng kiểm tra t để kiểm tra giả thuyết $H_0: a = 0$. Giá trị $t_{\text{tính toán}}$ được ước tính theo công thức [8.36].

$$t_a = \frac{a}{SE_a} \quad [8.36]$$

Trong đó giá trị của a được ước tính theo công thức [8.18]. SE_a được ước tính theo công thức [8.37]

$$SE_a = SE \sqrt{\frac{\sum_{i=1}^N X_i^2}{n \sum_{i=1}^N (X_i - \bar{X})^2}} \quad [8.37]$$

Giá trị t_a được so sánh với giá trị t_{tra} bảng với độ tự do df_E và mức ý nghĩa $\alpha/2$ (kiểm tra hai đuôi). Nếu $t_a \geq t_{\text{tra}}$ bảng, chúng ta bác bỏ H_0 có nghĩa hệ số a tồn tại (khác 0).

Bảng 8.5. Kết quả kiểm tra sự tồn tại hệ số a và b trong phương trình hồi quy ở ví dụ 8.11

Hệ số	Giá trị hệ số	SE	$t_{\text{tính toán}}$	P
a	6,59	2,29	2,87	0,008
b	4,40	0,77	5,71	0,000

Như vậy qua kết quả bảng 8.5 chúng ta bác bỏ giả thuyết $H_0: b = 0$ và giả thuyết $H_0: a = 0$.

8.3. Bài tập

- Giải thích mối quan hệ giữa các giá trị của x và y trong mẫu trong mỗi trường hợp hệ số tương quan mẫu như sau:
 - $r = 1$
 - $r = -1$
 - $r = 0,5$
 - $r = 0$
 - $r = -0,9$
 - $r = 0,77$

2. Kết quả xác định vòng ngực (bằng thước dây) và khối lượng (bằng cân điện tử) của 22 cá thể bò Lai Sind được thể hiện qua bảng sau. Hãy xây dựng phương trình hồi quy ước tính khối lượng từ vòng ngực?

Bảng 8.6. Vòng ngực và khối lượng của bò Lai Sind

Vòng ngực (cm)	Khối lượng (kg)
146	255
132	105,5
169	298
157,5	202
163	312
153	264
160	251
167	360
157	288
154	266
128	168
160	340
165	367
128	254
146	250
148	273
140	259
138	198,5
136	213
138	186
153	270
134	186

Nguồn: Lê Đình Phùng (2002)

3. Kết quả xác định khối lượng (kg), vòng ngực (cm), dài thân chéo (cm) và cao vây (cm) của cừu Phan Rang sơ sinh tại Ba Vì, Hà Nội được thể hiện ở bảng trang sau:

Bảng 8.7. Khối lượng, vòng ngực, dài thân chéo và cao vây của cừu Phan Rang sơ sinh nuôi tại Ba Vì, Hà Nội

Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)	Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)
1,9	28,0	27,0	28,0	2,0	26,0	25,0	26,0
2,2	27,0	29,0	28,0	2,1	26,0	25,0	28,0
2,3	28,0	26,0	31,0	2,2	26,0	28,0	26,0
2,0	27,0	27,0	28,0	2,0	27,0	27,0	29,0
2,0	29,0	27,0	29,0	2,1	28,0	29,0	28,0
2,1	26,0	25,0	29,0	1,9	28,0	27,0	28,0
2,3	25,0	28,0	28,0	2,3	28,0	26,0	31,0
2,0	27,0	27,0	28,0	2,1	26,0	25,0	28,0
2,1	27,0	28,0	27,0	2,4	29,0	29,0	32,0
2,2	28,0	27,0	30,0	1,9	28,0	27,0	28,0
2,3	26,0	28,0	28,0	2,4	29,0	29,0	30,0
2,0	27,0	27,0	28,0	2,3	28,0	26,0	31,0
2,1	29,0	30,0	30,0	2,2	28,0	29,0	30,0
2,4	30,0	28,0	29,0	2,3	28,0	26,0	31,0
2,1	27,0	29,0	26,0	2,5	30,0	32,0	33,0
2,4	31,0	31,0	32,0	1,9	28,0	30,0	29,0
2,0	27,0	27,0	28,0	2,2	28,0	29,0	28,0
2,0	29,0	27,0	29,0	2,3	26,0	29,0	29,0
2,5	30,0	32,0	33,0	1,9	28,0	27,0	28,0
2,1	26,0	25,0	28,0	2,2	27,0	29,0	28,0
2,6	28,0	28,0	31,0	2,4	29,0	29,0	30,0
2,1	26,0	25,0	28,0	2,1	26,0	25,0	28,0
2,0	29,0	27,0	29,0	2,2	27,0	29,0	28,0
2,5	28,0	30,0	29,0	1,9	28,0	27,0	28,0
1,9	28,0	27,0	28,0	2,5	27,0	26,0	31,0
2,2	25,0	28,0	29,0	2,1	28,0	29,0	28,0
2,0	26,0	25,0	26,0	2,0	26,0	25,0	26,0
2,5	29,0	26,0	30,0	2,3	26,0	29,0	29,0
2,4	30,0	29,0	31,0	2,4	29,0	29,0	30,0
2,3	27,0	28,0	31,0	2,0	27,0	27,0	29,0
2,5	27,0	26,0	31,0	2,1	28,0	29,0	28,0

Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)	Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)
2,4	27,0	28,0	29,0	2,0	27,0	27,0	29,0
2,6	30,0	29,0	33,0	2,0	27,0	27,0	29,0
2,1	26,0	25,0	28,0	2,5	27,0	26,0	31,0
2,2	27,0	29,0	28,0	2,3	26,0	28,0	28,0
2,2	26,0	28,0	26,0	2,6	30,0	30,0	31,0
1,9	28,0	27,0	28,0	2,5	28,0	30,0	29,0
2,3	26,0	29,0	29,0	2,3	26,0	28,0	28,0
2,4	30,0	28,0	29,0	2,4	30,0	29,0	31,0
2,3	25,0	28,0	28,0	2,9	31,0	32,0	33,0
2,4	30,0	28,0	29,0	2,5	28,0	30,0	29,0
2,0	27,0	27,0	28,0	2,2	28,0	27,0	30,0
2,2	26,0	28,0	26,0	2,3	27,0	28,0	31,0
2,4	30,0	28,0	29,0	2,3	27,0	29,0	28,0
2,3	25,0	28,0	28,0	2,5	27,0	28,0	29,0
2,5	28,0	30,0	29,0	2,1	28,0	29,0	28,0
2,0	27,0	27,0	28,0	2,4	27,0	28,0	29,0
1,7	26,0	23,0	26,0	2,3	27,0	28,0	31,0
2,6	29,0	27,0	34,0	2,4	27,0	28,0	29,0
2,1	28,0	29,0	28,0	2,3	27,0	28,0	31,0
2,3	25,0	28,0	28,0	2,4	27,0	28,0	29,0
2,4	31,0	31,0	32,0	2,1	26,0	25,0	29,0
2,0	27,0	27,0	28,0	2,3	27,0	29,0	28,0
1,8	24,0	25,0	28,0	2,2	25,0	28,0	29,0
2,3	26,0	28,0	28,0	2,4	29,0	29,0	32,0
2,4	31,0	31,0	32,0	2,3	27,0	29,0	28,0
2,2	26,0	28,0	26,0	2,0	29,0	27,0	29,0
2,4	31,0	31,0	32,0	2,3	27,0	29,0	28,0
2,4	30,0	29,0	31,0	2,4	29,0	29,0	32,0
2,2	28,0	27,0	30,0	1,9	25,0	26,0	27,0
2,8	29,0	27,0	33,0	2,2	25,0	28,0	29,0
2,4	30,0	29,0	31,0	2,4	29,0	29,0	32,0
2,2	28,0	27,0	30,0	2,6	30,0	29,0	33,0
2,3	26,0	29,0	29,0	2,0	29,0	27,0	29,0
2,2	28,0	29,0	30,0	2,3	27,0	28,0	31,0
2,0	29,0	27,0	29,0	1,9	25,0	26,0	27,0
2,4	29,0	29,0	30,0	2,0	29,0	27,0	29,0
2,1	29,0	29,0	30,0	1,7	26,0	23,0	26,0
2,4	29,0	29,0	30,0	2,2	27,0	29,0	28,0

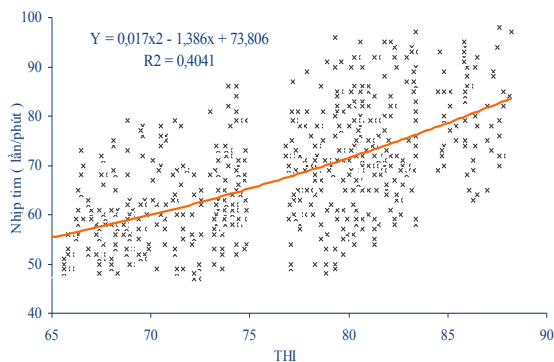
Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)	Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)
2,1	30,0	28,0	29,0	2,1	26,0	25,0	28,0
2,3	30,0	28,0	29,0	2,1	26,0	25,0	28,0
2,4	30,0	28,0	29,0	2,0	26,0	25,0	26,0
2,2	31,0	31,0	32,0	2,1	26,0	25,0	28,0
2,4	31,0	31,0	32,0	2,0	26,0	25,0	26,0
2,1	31,0	31,0	32,0	2,3	27,0	29,0	28,0
2,5	30,0	29,0	31,0	2,4	29,0	29,0	30,0
2,0	30,0	29,0	31,0	1,9	30,0	28,0	29,0
2,2	30,0	29,0	31,0	2,0	31,0	31,0	32,0
2,1	27,0	28,0	29,0	2,1	30,0	29,0	31,0
2,3	27,0	28,0	29,0	2,3	27,0	28,0	29,0
2,3	27,0	28,0	29,0	2,4	29,0	29,0	32,0
2,4	29,0	29,0	32,0	2,1	29,0	29,0	30,0
2,1	29,0	29,0	32,0	2,1	29,0	29,0	30,0
2,1	29,0	29,0	32,0	2,0	29,0	29,0	30,0
1,8	26,0	26,0	27,0	2,1	30,0	28,0	29,0
2,1	29,0	30,0	30,0	2,4	30,0	28,0	29,0
2,3	25,0	28,0	28,0	2,1	30,0	28,0	29,0
2,2	28,0	29,0	28,0	2,4	31,0	31,0	32,0
2,5	27,0	26,0	31,0	2,0	31,0	31,0	32,0
2,05	27	28	29	2,3	31,0	31,0	32,0
2,3	25,0	28,0	28,0	2,1	30,0	29,0	31,0
2,3	26,0	28,0	28,0	2,0	30,0	29,0	31,0
2,3	30,0	29,0	31,0	2,4	30,0	29,0	31,0
2,4	27,0	28,0	29,0	2,4	30,0	29,0	31,0
2,6	27,0	28,0	29,0	2,0	30,0	29,0	31,0
2,3	27,0	28,0	29,0	2,0	27,0	28,0	29,0
2,1	29,0	29,0	32,0	2,2	27,0	28,0	29,0
2,3	29,0	29,0	32,0	2,3	27,0	28,0	29,0
2,4	29,0	29,0	32,0	2,0	29,0	29,0	32,0
1,7	29,0	29,0	30,0	2,4	29,0	29,0	32,0
2,1	29,0	29,0	30,0	2,0	29,0	29,0	32,0
2,3	30,0	28,0	29,0	2,3	29,0	29,0	30,0
2,1	31,0	31,0	32,0	2,4	29,0	29,0	30,0
2,4	29,0	29,0	32,0	2,1	30,0	28,0	29,0
2,8	29,0	28,0	33,0	2,0	31,0	31,0	32,0
2,1	26,0	25,0	28,0	2,3	29,0	29,0	32,0
2,4	29,0	29,0	30,0	2,2	28,0	27,0	30,0

Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)	Khối lượng (kg/con)	Vòng ngực (cm)	Dài thân chéo (cm)	Cao vây (cm)
2,0	30,0	28,0	29,0	2,2	28,0	27,0	30,0
2,0	31,0	31,0	32,0	2,0	27,0	27,0	28,0
2,4	30,0	29,0	31,0	2,1	28,0	29,0	28,0
2,5	27,0	28,0	29,0	2,4	29,0	29,0	30,0
2,0	29,0	29,0	32,0	2,4	30,0	28,0	29,0
2,2	29,0	29,0	30,0	2,4	31,0	31,0	32,0
2,2	29,0	29,0	30,0	2,5	30,0	29,0	31,0
2,3	29,0	29,0	30,0	2,0	27,0	28,0	29,0
2,1	30,0	28,0	29,0	2,2	29,0	29,0	32,0
2,3	30,0	28,0	29,0	1,9	29,0	29,0	30,0
2,0	30,0	28,0	29,0	2,0	29,0	29,0	30,0
2,1	31,0	31,0	32,0	2,4	29,0	29,0	30,0

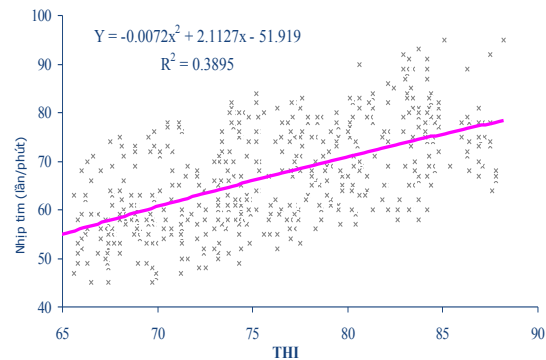
Nguồn: Nguyễn Thành Vinh (2010)

- Hãy xây dựng biểu đồ phân tán thể hiện mối quan hệ giữa khối lượng với vòng ngực, dài thân chéo với cao vây? Từ biểu đồ phân tán, hãy nhận xét về sức mạnh của mối quan hệ giữa khối lượng với các chiều đo?
- Hãy ước tính và kiểm tra sự tồn tại của hệ số tương quan tuyến tính giữa khối lượng với vòng ngực, dài thân chéo với cao vây?
- Xây dựng phương trình hồi quy tuyến tính ước tính khối lượng từ vòng ngực, dài thân chéo và cao vây; kiểm tra sự tồn tại của từng phương trình hồi quy, kiểm tra sự tồn tại của từng tham số trong phương trình hồi quy và ước tính giá trị của hệ số quyết định có hiệu chỉnh của từng phương trình hồi quy?

Đồ thị 8.11 và 8.12 thể hiện tương quan hồi quy giữa nhịp tim của cừu Phan Rang 6 - 12 tháng tuổi và trên 12 tháng tuổi nuôi tại Thừa Thiên Huế với chỉ số nhiệt ẩm (THI). Các tác giả đã sử dụng hàm hồi quy bậc 2 để hệ thể hiện mối quan hệ giữa nhịp tim và THI. Hãy bình luận kết quả của các tác giả ?



Đồ thị 8.11: Tương quan giữa THI với nhịp tim của cừu 6 - 12 tháng tuổi



Đồ thị 8.12: Tương quan giữa THI với nhịp tim của cừu trên 12 tháng tuổi

Nguồn: Bùi Văn Lợi và cộng sự (2010)

4. Một nhóm nghiên cứu ảnh hưởng của lượng protein ăn vào đến tăng trọng của bò Vàng Việt Nam trong giai đoạn vỗ béo. Nghiên cứu được tiến hành trên 24 cá thể bò; kết quả được thể hiện ở bảng 8.8

Bảng 8.8. Lượng protein ăn vào và tăng trọng của bò Vàng Việt Nam trong giai đoạn vỗ béo

Protein ăn vào (kg/ngày)	Tăng trọng (g/ngày)	Protein ăn vào (kg/ngày)	Tăng trọng (g/ngày)
0,35	331	0,53	502
0,47	649	0,56	739
0,45	547	0,60	667
0,47	651	0,63	818
0,48	662	0,58	505
0,51	655	0,70	833
0,55	691	0,66	520
0,60	829	0,75	806
0,73	775	0,78	601
0,78	716	0,82	865
0,71	682	0,87	986
0,78	815	0,89	941

Nguồn: Đinh Văn Dũng và cộng sự (2009)

- a. Hãy xây dựng biểu đồ phân tán thể hiện mối quan hệ giữa lượng protein ăn vào và tăng trọng của bò Vàng Việt Nam? Từ biểu đồ phân tán hãy suy diễn về mối quan hệ tuyến tính giữa 2 biến nghiên cứu này?
- b. Hãy xác định hệ số tương quan tuyến tính và xác định sự tồn tại của hệ số tương quan tuyến tính
- c. Xác định phương trình hồi quy tuyến tính giữa lượng protein ăn vào và tăng trọng của bò? Kiểm tra sự tồn tại của các hệ số trong phương trình hồi quy?
- d. Xác định hệ số quyết định giải thích ảnh hưởng của lượng protein ăn vào đến tăng trọng của bò Vàng Việt Nam trong giai đoạn vỗ béo?

Chương 9

NGHIÊN CỨU ĐIỀU TRA BẰNG BẢN HỎI, BẢN KIỂM

Chương này đề cập đến một phương pháp nghiên cứu khác với các phương pháp nghiên cứu đã đề cập trong trước đây, đó là nghiên cứu điều tra bằng bản hỏi. Các nội dung chính sẽ được trình bày bao gồm: vị trí của nghiên cứu điều tra bằng bản hỏi trong hệ thống các phương pháp nghiên cứu và các bước để tiến hành một nghiên cứu điều tra bằng bản hỏi.

9.1. Khái niệm

Nghiên cứu điều tra bằng bản hỏi là nghiên cứu trong đó các thông tin được thu thập thông qua một tập hợp các loại câu hỏi khác nhau được chuẩn bị trước khi tiến hành nghiên cứu chính thức.

Một khái niệm chúng ta có thể gặp đó là bản kiểm. Bản hỏi hay bản kiểm đều là công cụ để thu thập thông tin trong nghiên cứu điều tra. Bản hỏi thu thập thông tin trực tiếp qua các câu hỏi trong khi đó bản kiểm thu thập thông tin gián tiếp qua các câu hỏi. Bản hỏi là tập hợp các câu hỏi để lấy thông tin, được sắp xếp một cách logic. Người sử dụng bản hỏi để thu thập thông tin có thể không cần hiểu một cách chi tiết về mục tiêu, nội dung nghiên cứu mà chỉ lần lượt đặt từng câu hỏi đã được chuẩn bị trước và điền thông tin vào bản hỏi.

Bản kiểm là biểu mẫu để thu thập các thông tin. Cách khai thác thông tin có thể linh động tùy theo người phỏng vấn và đối tượng bị phỏng vấn. Khi người điều tra cũng là người thiết kế nghiên cứu hay là thành viên của nhóm nghiên cứu thì nên sử dụng bản kiểm. Khi người điều tra không phải là người thiết kế nghiên cứu hay thành viên nhóm nghiên cứu thì nên sử dụng bản hỏi. Linh động trong việc sử dụng bản hỏi hay bản kiểm trong nghiên cứu điều tra nhằm nâng cao hiệu quả nhập và quản lý số liệu nghiên cứu.

Số bò bê tại thời điểm hiện tại (con):

Loại bò	Nội (#)		Lai (#)	
	Đực	Cái	Đực	Cái
< 6 tháng				
7-12 tháng				
13-18 tháng				
19-24 tháng				
> 24 tháng				
Trong số bò lớn hơn 24 tháng				
Bò sinh sản				
Bò cày kéo				
Bò vỗ béo thâm canh				
Số bò nuôi vỗ béo thâm canh				

Biểu đồ 9.1. Ví dụ trích đoạn của bản kiểm (Nguồn: Lê Đình Phùng, 2007)

1. Gia đình Bác có bao nhiêu con bò Vàng dưới 6 năm tuổi?
 - a. Trong đó có bao nhiêu con cái?
 - b. Trong đó có bao nhiêu con đực?
2. Gia đình bác có nuôi bò lai không? ☐ Không Nếu có tiếp câu 3
3. Gia đình bác có nuôi bao nhiêu con đực lai?
4. ...

Biểu đồ 9.2. Ví dụ trích đoạn của bản hỏi (Nguồn: Lê Đình Phùng, 2007)

Trong phạm vi của chương này chúng tôi sẽ chủ yếu lấy các ví dụ về bản kiểm.

9.2. Vị trí của nghiên cứu điều tra bằng bản hỏi trong hệ thống các nghiên cứu

Như đã đề cập trong các chương trước, có ít nhất 8 quan điểm phân loại phương pháp nghiên cứu khác nhau, một trong những quan điểm đó là chia nghiên cứu làm hai loại: nghiên cứu điều tra và nghiên cứu thí nghiệm.

Nghiên cứu thí nghiệm là các loại nghiên cứu có thay đổi bản chất của biến độc lập để quan sát sự thay đổi của biến phụ thuộc. Ví dụ, điều chỉnh mức protein trong khẩu phần để quan sát sự thay đổi về khả năng tăng trọng, hay bổ sung hay không bổ sung DL-Methionin hay không bổ sung vào khẩu phần để quan sát sự thay đổi hiệu quả chuyển hóa thức ăn. Các loại nghiên cứu thí nghiệm cơ bản đã được đề cập trong chương 7.

Nghiên cứu điều tra là nghiên cứu phản ánh bản chất của sự vật hiện tượng mà không có sự tác động đến sự vật hiện tượng, hay không có sự thay đổi biến độc lập để quan sát sự thay đổi của biến phụ thuộc. Điều này không có nghĩa nghiên cứu điều tra không có biến độc lập. Nghiên cứu điều tra hoàn toàn có thể có biến độc lập. Tuy nhiên,

cụm từ ảnh hưởng của biến độc lập lên biến phụ thuộc rất ít khi được dùng trong nghiên cứu điều tra. Thông thường cụm từ sự sai khác của biến phụ thuộc giữa các mức của biến độc lập được sử dụng. Ví dụ thông thường chúng ta nói “Ảnh hưởng của mức protein trong khẩu phần đến tăng trọng của lợn lai F1 (Landrace × Yorkshire....” và “Sự sai khác về khả năng tăng trọng của lợn lai F1 (Landrace × Yorkshire) ở vùng sinh thái đồng bằng và miền núi...”.

Nghiên cứu điều tra gồm có hai loại chính: Nghiên cứu quan sát và nghiên cứu điều tra bằng bản hỏi. Nghiên cứu điều tra bằng bản hỏi gồm có hai nhóm chính: Nghiên cứu điều tra bằng bản hỏi tại một thời điểm (không lặp lại) và nghiên cứu điều tra bằng bản hỏi có lặp lại. Trong phạm vi của một đề tài nghiên cứu thuộc bậc đại học và thạc sĩ, hay các đề tài không mang tính dài hạn thông thường chúng ta tiến hành nghiên cứu điều tra không lặp lại (cross sectional survey). Ví dụ điều tra tình hình chăn nuôi bò tại một xã A. Các chỉ tiêu nghiên cứu bao gồm quy mô, cơ cấu đàn theo giống, cơ cấu đàn theo tuổi, hiệu quả chăn nuôi... Các kết quả nghiên cứu phản ánh số liệu tại một thời điểm đó là thời điểm điều tra.

Trong các nghiên cứu dài hạn như nghiên cứu sinh bậc tiến sĩ, đặc biệt là các nghiên cứu muốn quan sát sự thay đổi theo thời gian, người ta tiến hành nghiên cứu điều tra lặp lại (longitudinal survey). Ví dụ nghiên cứu điều tra về khả năng sinh trưởng theo thời gian của tổ hợp lai giữa Simental và Lai Sind trong điều kiện chăn nuôi nông hộ tại một xã A nào đó. Chúng ta có thể đạt được mục tiêu này bằng cách điều tra tại một thời điểm trên các đối tượng bò có lứa tuổi khác nhau. Chúng ta cũng có thể đạt được mục tiêu này bằng cách tiến hành điều tra trên một số đối tượng bò cố định, nhưng tiến hành điều tra lặp lại theo thời gian. Cách này còn gọi là nghiên cứu điều tra lặp lại.

9.3. Các bước trong tiến hành nghiên cứu điều tra bằng bản hỏi

Sau khi đã xác định được vấn đề nghiên cứu, nghiên cứu điều tra bằng bản hỏi gồm có các bước cơ bản sau:

1. Xác định mục tiêu nghiên cứu
2. Xác định giả thuyết nghiên cứu
3. Xác định nhân tố nghiên cứu (nếu có)
4. Xác định chỉ số nghiên cứu
5. Xác định đối tượng và mức độ nghiên cứu
6. Lượng hóa các chỉ số
7. Xây dựng mô hình lý thuyết
8. Phát triển bản hỏi
9. Thử nghiệm bản hỏi
10. Thu thập số liệu
11. Nhập số liệu
12. Phân tích số liệu
13. Báo cáo kết quả

9.3.1. Xác định mục tiêu nghiên cứu

Cần phân biệt giữa mục đích và mục tiêu nghiên cứu. Mục đích nghiên cứu để trả lời câu hỏi tiến hành nghiên cứu này để làm gì? Trong khi đó mục tiêu nghiên cứu cụ thể hơn, được dùng để trả lời câu hỏi nghiên cứu này làm cái gì? Mục tiêu nghiên cứu cần được xác định rõ ràng, cụ thể, đo lường được và có thể đạt được (tiêu chuẩn SMART). Khi xác định mục tiêu nghiên cứu cần phải chú ý nghiên cứu xuất phát từ nhu cầu nào? Đã có công trình nào được thực hiện chưa? Nếu đã được thực hiện thì lỗ hổng nghiên cứu là ở đâu?

Trong nghiên cứu điều tra bằng bản hỏi, khi mục tiêu nghiên cứu chưa rõ ràng thì không nên tiến hành các bước tiếp theo của nghiên cứu, bởi vì các bước nghiên cứu tiếp theo hoàn toàn phụ thuộc vào mục tiêu nghiên cứu, đặc biệt là bước xác định chỉ số nghiên cứu. Một số mục tiêu nghiên cứu ví dụ như sau:

1. Đánh giá khả năng sinh sản của lợn nái Móng Cái nuôi trong nông hộ tại huyện A tỉnh B.
2. Đánh giá sự sai khác về khả năng và hiệu quả chăn nuôi bò lai F1 (Vàng × Brahman) tại vùng sinh thái miền núi và đồng bằng của tỉnh A.
3. Đánh giá hệ thống chăm sóc nuôi dưỡng bò tại huyện A.

Trong các đề tài điều tra bằng bản hỏi đánh giá sự sai khác giữa các mức của biến độc lập, ví dụ giữa các vùng sinh thái khác nhau, giữa các hệ thống chăn nuôi khác nhau, mục tiêu nghiên cứu nên được xác định theo cách thể hiện cả biến độc lập và biến phụ thuộc.

9.3.2. Xác định giả thuyết nghiên cứu

Giả thuyết cần được nêu rõ ràng, có thể kiểm tra được, thể hiện mối quan hệ giữa biến độc lập và biến phụ thuộc và có phạm vi nhất định. Có thể hiểu giả thuyết là kết luận sơ bộ của đề tài nghiên cứu, hay là câu trả lời cho câu hỏi nghiên cứu hoặc là câu trả lời dự đoán.

Có hai loại giả thuyết: Giả thuyết không khuynh hướng và giả thuyết có khuynh hướng. Giả thuyết không khuynh hướng sử dụng các cụm từ như “ảnh hưởng”, “sai khác”. Loại giả thuyết này không chỉ ra khuynh hướng của sự sai khác. Giả thuyết này còn gọi là giả thuyết hai đuôi (liên quan đến sự bố trí trên hai đuôi). Giả thuyết có khuynh hướng được nêu ra như một kết quả dự kiến. Loại giả thuyết này thông thường sử dụng các cụm từ như *giảm, tăng, thấp hơn, cao hơn* - Giả thuyết có khuynh hướng còn được gọi là giả thuyết một đuôi (bên phải hoặc bên trái).

9.3.3. Xác định nhân tố nghiên cứu

Nhân tố nghiên cứu là nguyên nhân ảnh hưởng đến các đặc điểm quan sát. Trong ví dụ nêu trên nhân tố nghiên cứu là giống/tổ hợp lai bò, vùng sinh thái. Nhân tố bao gồm

các mức của nhân tố. Ví dụ nhân tố giống/tổ hợp lai bò nêu trên gồm có hai mức đó là giống bò Vàng và tổ hợp lai F1 (Vàng $\times$ Brahman). Nhân tố có thể là định tính hoặc/và định lượng. Ví dụ nhân tố định tính là các tổ hợp bò lai và nhân tố định lượng là các mức protein thô trong khẩu phần thức ăn cho bò.

Xác định nhân tố là điều kiện tiên quyết để thiết kế bản hỏi, đồng thời thiết kế mã hóa số liệu, đặc biệt xác định nhân tố nghiên cứu quyết định phương án phân tích số liệu. Trong nghiên cứu điều tra cần thiết phải có phương án phân tích số liệu trước khi thiết kế bản hỏi. Khi tất cả các nhân tố nghiên cứu là định lượng thì phương án phân tích số liệu cơ bản là phân tích hồi quy. Khi tất cả nhân tố nghiên cứu là định tính thì phương án phân tích số liệu cơ bản là phân tích phương sai và khi nhân tố nghiên cứu vừa định tính và vừa định lượng thì dùng mô hình hỗn hợp để phân tích số liệu điều tra.

Nhân tố nghiên cứu do chính nhà nghiên cứu quyết định. Không phải nghiên cứu nào có nhân tố nghiên cứu. Những nghiên cứu mô tả sự vật hiện tượng thông thường không có nhân tố nghiên cứu.

9.3.4. Xác định chỉ số nghiên cứu

Chỉ số nghiên cứu có thể được xem là biến phụ thuộc. Thông thường chỉ số được phân loại dựa trên tính chất và dựa trên nghiên cứu tác động. Dựa trên tính chất, chỉ số được phân ra

- Chỉ số định lượng (Số lượng, tần xuất, thời gian)
- Chỉ số định tính (ai, khi nào, cái gì, thế nào)

Dựa trên nghiên cứu tác động, chỉ số được phân ra

- Chỉ số kết quả (dài hạn)
- Chỉ số tác động (ngắn, trung hạn)
- Chỉ số quá trình (theo các hoạt động can thiệp)
- Chỉ số tiến trình (Đo chỉ số theo từng giai đoạn)

Để có thể xác định được chỉ số nghiên cứu, cần căn cứ vào mục tiêu nghiên cứu và căn cứ vào khả năng thu thập thông tin. Giống như xác định mục tiêu chỉ số nghiên cứu cũng cần phải SMART.

Ví dụ 9.1. Một nhóm nghiên cứu quan tâm đến sự sai khác nhau về hiệu quả chăn nuôi bò Vàng và bò lai F1 (Vàng $\times$ Brahman) ở vùng sinh thái đồng bằng và miền núi của tỉnh Bình Định.

Như vậy:

- Nhân tố nghiên cứu là: Giống/Tổ hợp lai; Vùng sinh thái
- Chỉ số nghiên cứu là: Hiệu quả chăn nuôi bò

Nhóm nghiên cứu cần xác định rõ hiệu quả chăn nuôi bò trên góc độ nào? Kinh tế, xã hội hay môi trường. Giả sử hiệu quả kinh tế được nghiên cứu. Nhóm nghiên cứu cần cụ thể hơn hay đặt câu hỏi chỉ số nào sẽ được dùng để đánh giá hiệu quả kinh tế. Ví dụ một số chỉ số được sử dụng như:

- Tổng thu từ chăn nuôi bò của hộ
- Chi phí cho chăn nuôi bò
- Thu nhập cận biên từ chăn nuôi bò
- Hiệu quả đầu tư đồng vốn cho chăn nuôi bò

Ví dụ 9.2. Một nhóm nghiên cứu quan tâm đến sự sai khác giữa các tổ hợp bò lai F₁ (Vàng × Brahman), F₁ (Vàng × Brahman) × Droughtmaster, F₁ (Vàng × Brahman) × Limousin về khả năng sản xuất khi được nuôi trong hai hệ thống chăn nuôi khác nhau (Bán thâm canh và Thâm canh).

Như vậy:

- Nhân tố nghiên cứu là: Tổ hợp bò lai; Hệ thống chăn nuôi
- Chỉ số nghiên cứu là: Khả năng sản xuất

Nhóm nghiên cứu cần xác định các chỉ số đánh giá khả năng sản xuất của bò, hay cụ thể chính là các tính trạng đánh giá khả năng sản xuất của bò như vòng ngực, dài thân chéo, tăng trọng...

Ví dụ 9.3. Một nhóm nghiên cứu quan tâm đến hiện trạng chăn nuôi bò tại xã A

Trong trường hợp này nhóm nghiên cứu cần xác định các chỉ số đánh giá hiện trạng chăn nuôi, như quy mô, cơ cấu giống, cơ cấu tuổi, cơ cấu theo mục tiêu sử dụng...

9.3.5. Lượng hóa chỉ số nghiên cứu

Chúng ta cần xác định cách ước tính (lượng hóa) từng chỉ số một. Để có thể lượng hóa được các chỉ số chúng ta cần căn cứ vào kiến thức chuyên môn liên quan đến chỉ số, căn cứ vào nguồn thông tin có sẵn và căn cứ vào điều kiện thực tế.

Trong ví dụ 9.1 nêu trên, chỉ số nghiên cứu hiệu quả chăn nuôi bò được xác định như sau:

- Hiệu quả = (Thu nhập cận biên - Chi phí biến động)/Chi phí biến động
- Thu nhập cận biên = Tổng giá trị đầu ra - Chi phí biến động
- Chi phí biến động = Tổng các chi phí diễn ra trong năm như thức ăn, thuốc thú y, thuê nhân công lao động.
- Tổng giá trị đầu ra =

Quá trình xác định các chỉ số thường diễn ra cho đến khi các thông tin có thể được khai thác trực tiếp trên đối tượng cung cấp thông tin.

9.3.6. Xây dựng mô hình định tính

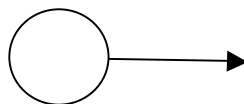
Trong nghiên cứu điều tra bằng bản hỏi, mô hình định tính chính là mô hình hệ thống hay chính là mô hình sử dụng các công cụ mô phỏng các yếu tố của hệ thống để mô phỏng đối tượng điều tra và các yếu tố có liên quan đến đối tượng điều tra.

Mô hình hệ thống hình ảnh hóa tất cả đầu vào, đầu ra, mối quan hệ, các yếu tố ảnh hưởng đến đối tượng nghiên cứu. Mô hình hệ thống bao gồm các thông tin cần thiết để lượng hóa các chỉ số nghiên cứu. Để xây dựng được mô hình định tính cần thiết phải xác định được đối tượng nghiên cứu, ví dụ vật nuôi, nông hộ hay cộng đồng.

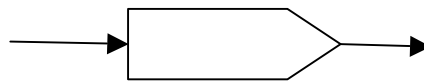
Mục đích của việc xây dựng mô hình định tính là để giúp cho việc soạn bản hỏi một cách dễ dàng và điều quan trọng hơn là không bỏ sót thông tin. Rất nhiều nghiên cứu điều tra, sau khi kết thúc điều tra mới phát hiện được rằng bản hỏi thiếu một số câu hỏi để thu thập đầy đủ thông tin phục vụ cho nghiên cứu. Xây dựng mô hình định tính càng quan trọng khi chúng ta càng có nhiều chỉ số nghiên cứu. Trong thực tế rất nhiều nghiên cứu không xây dựng mô hình định tính để làm cơ sở cho thiết kế bản hỏi cho nên đã không có một bản hỏi tối ưu.

Nhóm nghiên cứu có thể tự dùng các ngôn ngữ (hình ảnh) riêng của mình để xây dựng mô hình định tính miễn là các thành viên trong nhóm nghiên cứu hiểu được ngôn ngữ đó. Tuy nhiên, để có thể xuất bản kết quả nghiên cứu trên các tạp chí khoa học thì cần phải sử dụng các ngôn ngữ chuẩn. Một số ví dụ như sau:

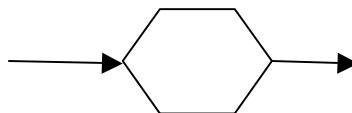
- Nguồn lực từ bên ngoài



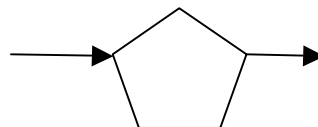
- Hệ thống phụ sản xuất



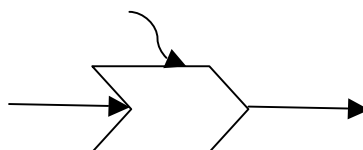
- Hệ thống phụ tiêu thụ



- Hệ thống phụ dự trữ thụ động



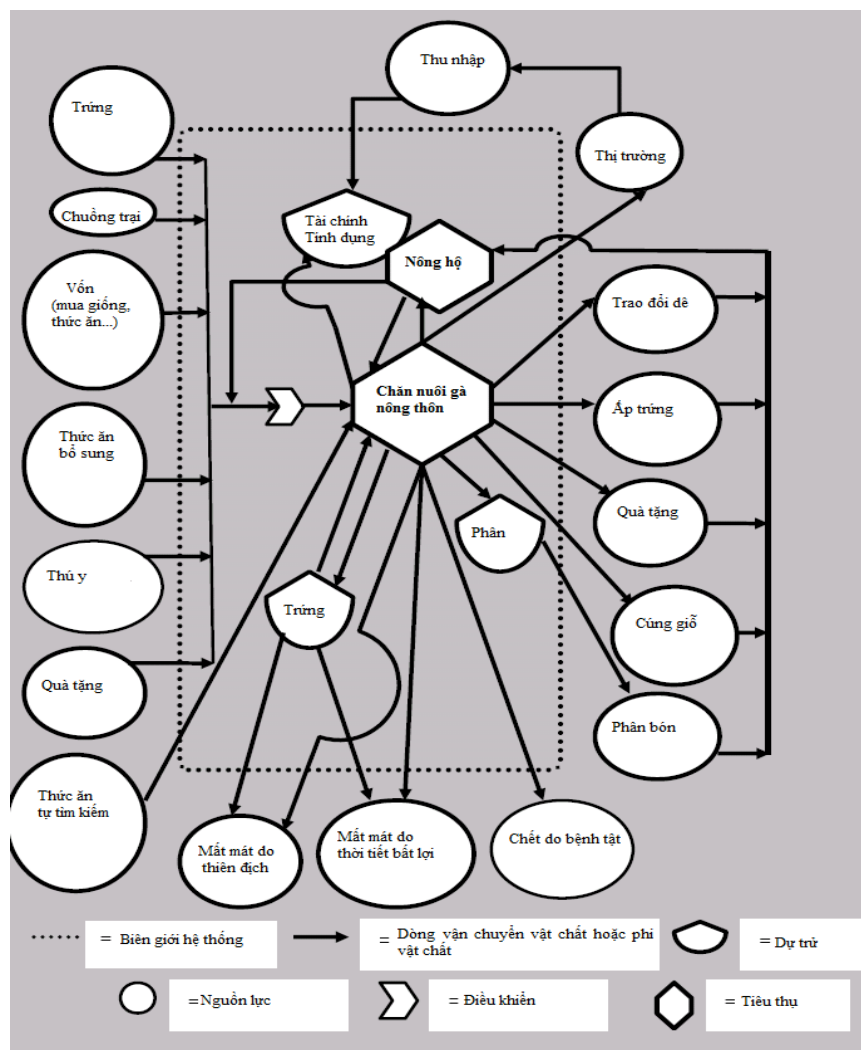
- Điều khiển quản lý



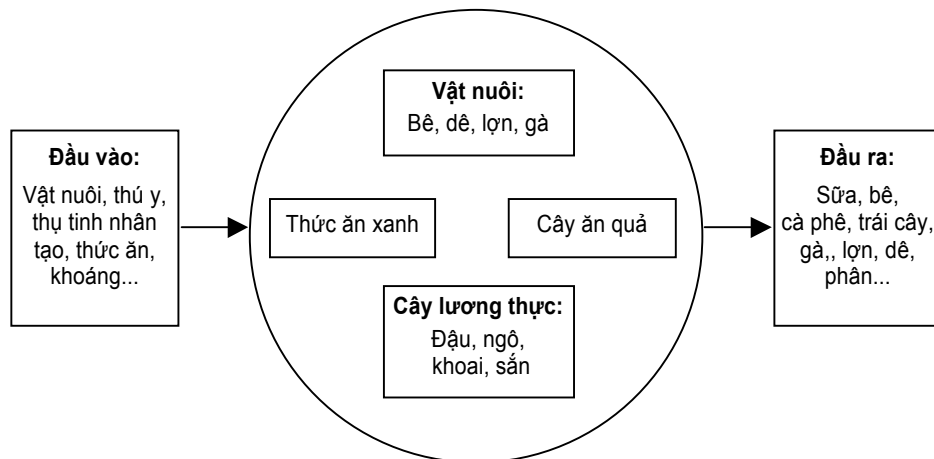
Các ký hiệu này được dùng trong lĩnh vực mô hình hóa và nhiều phần mềm tin học khác nhau có thể nhận biết các ký hiệu này ví dụ phần mềm Stella.

Ví dụ trong biểu đồ 9.3 nêu ở dưới là mô hình định tính mô tả hệ thống chăn nuôi gà quy mô nhỏ ở Bukinafaxo. Mô hình xác định tất cả các yếu tố đầu vào, đầu ra của hệ thống cũng như mối quan hệ giữa các hệ thống phụ. Sau khi có mô hình định tính thì chúng ta có thể căn cứ vào mô hình định tính để thiết kế bản hỏi nghiên cứu.

Như đã đề cập ở trên, nhóm nghiên cứu có thể sử dụng ngôn ngữ riêng của nhóm mình để xây dựng mô hình định tính (biểu đồ 9.4). Mô hình mô tả hệ thống chăn nuôi bò sữa quy mô nhỏ vùng ven thành phố. Mô hình xác định các yếu tố đầu vào và đầu ra của hệ thống cũng như phản ánh mối quan hệ của các hệ thống phụ: cây trồng, vật nuôi, cây ăn quả và cây thức ăn.



Biểu đồ 9.3. Mô hình hệ thống chăn nuôi gà quy mô nhỏ ở Bukinafaxo (Kondombo, 2005)



Biểu đồ 9.4. Mô hình hệ thống chăn nuôi bò sữa quy mô nhỏ ven thành phố (Fonteh et al., 2006)

9.3.7. Phát triển bản hỏi

Sau khi đã hoàn thành các bước nêu trên, nhóm nghiên cứu có thể tiến hành thiết kế bản hỏi/bảng kiểm. Quá trình soạn bản hỏi là quá trình chuyển mục tiêu nghiên cứu sang dạng câu hỏi trong bản hỏi/bảng kiểm. Nguyên tắc của thiết kế bản hỏi là “Luôn luôn làm cho bản hỏi đơn giản” hay còn gọi là nguyên tắc KISS.

Một số bí quyết khi soạn bản hỏi như sau:

1. Bản hỏi cần được hiểu theo một nghĩa duy nhất, duy nhất theo từng từ, từng câu
2. Tránh tất cả các câu hỏi không rõ ràng
3. Tránh tất cả các câu hỏi đa nghĩa
4. Người trả lời có thể trả lời được các câu hỏi
5. Các câu hỏi nên trực tiếp đi đến vấn đề cần hỏi
6. Không nên đặt nhiều ý cần hỏi vào trong một câu hỏi
7. Tránh câu hỏi không xác định được ý nghĩa của câu trả lời.

Ví dụ 9.3. Câu hỏi “Có phải Ông/bà nuôi bò lai F₁ (Vàng × Brahman) và bò Vàng”

Câu trả lời: Vâng

Người nghiên cứu sẽ không biết được một cách chắc chắn ý nghĩa của “Vâng”. “Vâng” có thể là:

- ✓ Chỉ nuôi bò vàng
- ✓ Chỉ nuôi bò lai
- ✓ Hay nuôi cả hai loại bò

Câu hỏi “Ông/bà đã bao giờ trồng cỏ nuôi bò và nếu có thì Ông/bà có thích nó không?”

Câu trả lời: Không

Người nghiên cứu sẽ không biết được một cách chắc chắn về ý nghĩa của “Không”. “Không” có thể là:

- ✓ Chưa trồng
 - ✓ Và cũng có thể là đã trồng mà không thích
8. Bản hỏi cần được thiết kế với cách suy nghĩ về hướng phân tích số liệu thường trực trong đầu. Người thiết kế phải tự trả lời câu hỏi “nếu thu được thông tin từ câu hỏi này thì sẽ xử lý số liệu như thế nào?”
9. Khi soạn bản hỏi phải luôn đặt câu hỏi “sẽ làm gì với thông tin thu được từ mỗi câu hỏi”
10. Không nên đưa thêm một số câu hỏi không có mục đích vào trong bản hỏi
11. Nên xem xét và phân loại mỗi câu hỏi vào ba nhóm sau:
- ✓ Câu hỏi cần phải có thông tin
 - ✓ Câu hỏi sẽ rất hữu ích để có thông tin
 - ✓ Câu hỏi có thông tin thì cũng tốt
- Nên loại bỏ các câu hỏi nhóm cuối cùng nếu hai nhóm trước đã có câu hỏi đủ dài
12. Có thể bắt đầu bản hỏi bởi một đoạn giới thiệu về bản hỏi để người được hỏi hiểu được mục tiêu của cuộc điều tra. Đây là điều quan trọng đảm bảo độ chính xác của thông tin thu được.
13. Khi soạn bản hỏi có thể sử dụng 3 loại câu hỏi cơ bản

- Nhiều lựa chọn

Theo Ông/bà, nguồn thu nhập nào là quan trọng nhất của hộ gia đình?

- ☐ Chăn nuôi
- ☐ Trồng trọt
- ☐ Thủy sản
- ☐ Dịch vụ

- Câu hỏi định lượng mở

Gia đình Ông/bà chi bao nhiêu tiền thú y cho chăn nuôi bò/năm?

- Câu hỏi định tính mở

Theo Ông/bà làm thế nào để cải thiện chất lượng môi trường chăn nuôi vùng ven thành phố?

- Câu hỏi đánh giá mức độ

Theo Ông/bà sự thay đổi trữ lượng cỏ tự nhiên cho chăn nuôi bò như thế nào?

- ☐ Không tăng
- ☐ Giảm nhẹ

- ☐ Giảm tương đối
- ☐ Giảm nghiêm trọng

- Mức độ đồng ý

Ông/bà thích nuôi bò lai F_1 (Vàng $\times$ Brahman) ở mức độ nào?

- ☐ Rất thích
- ☐ Không thích lắm
- ☐ Hoàn toàn không thích

14. Trật tự câu hỏi và trật tự trả lời ảnh hưởng rất lớn đến kết quả điều tra, do vậy những câu hỏi đầu thường là câu hỏi dễ, tạo thích thú, lấy lòng người trả lời và những câu hỏi cuối cùng thường là những câu hỏi nhạy cảm.

15. Tối đa hóa việc sử dụng các câu hỏi nên đơn giản

Sau khi hoàn thành bản hỏi, chúng ta cần kiểm tra lại bản hỏi căn cứ vào các câu hỏi sau:

- *Chúng ta đã tránh tất cả các câu hỏi định hướng chưa?* Các câu hỏi không nên có các cụm từ định hướng như “ông/bà không nói rằng.....” hay “ông/bà không đồng ý rằng.....”
- *Câu hỏi đã cụ thể một cách tối đa chưa?* Các câu hỏi nên tránh dùng các từ như “đôi khi”, “thỉnh thoảng”, “thông thường”, “trong vùng này” bởi vì các cụm từ đó được hiểu theo một nghĩa khác nhau ở những người trả lời khác nhau.
- Các câu hỏi có thể được hiểu bởi tất cả các đối tượng phỏng vấn chưa?
- Mỗi một câu hỏi sẽ được áp dụng cho tất cả người trả lời chưa?

9.3.8. Thử nghiệm bản hỏi

Sau khi thiết kế bản hỏi, bản hỏi cần được kiểm tra (validate) trước khi được chính thức sử dụng để thu thập thông tin. Bản hỏi được thử nghiệm trên một số đối tượng nghiên cứu đại diện cho quần thể nghiên cứu. Lý tưởng nhất là bản hỏi được kiểm tra trên một nhóm đối tượng nghiên cứu tương đồng nhóm đối tượng sẽ được tiến hành nghiên cứu chính thức.

Thử nghiệm bản hỏi gồm một số mục tiêu sau:

1. Xem xét tính thực tế của các phép và đơn vị đo lường sử dụng trong bản hỏi
2. Giúp người điều tra hiểu hệ thống sản xuất, tạo điều kiện thuận lợi cho việc thu thập thông tin chính xác. Ví dụ khi người điều tra có đầy đủ thông tin về các loại thức ăn sử dụng cho bò ở xã A thì sẽ dễ dàng cho việc thu thập thông tin liên quan đến chế độ nuôi dưỡng bò trong nông hộ ở xã A.
3. Kiểm tra thời gian tiêu tốn để hoàn thành điều tra một bản hỏi. Thông tin này giúp cho việc lập kế hoạch điều tra một cách chính xác.
4. Kiểm tra câu hỏi để tránh những câu hỏi có nghĩa không rõ ràng

5. Kiểm tra và loại bỏ những câu hỏi mà không đem lại thông tin hữu ích

Thông thường bản hỏi được thử nghiệm trên số mẫu chiếm khoảng 5-10% tổng số mẫu nghiên cứu chính thức. Kết quả thu được từ kiểm tra bản hỏi không được đưa vào kết quả nghiên cứu cuối cùng. Mọi kỹ thuật kiểm tra bản hỏi cần được tiến hành giống như khi điều tra chính thức. Sau khi thử nghiệm bản hỏi, nhóm nghiên cứu cần điều chỉnh bản hỏi trước khi tiến hành điều tra chính thức. Thử nghiệm bản hỏi không gây tốn nhiều thời gian nhưng lại có ý nghĩa rất lớn trong việc đảm bảo hiệu quả điều tra cũng như độ chính xác của thông tin.

9.3.9. Điều tra chính thức

9.3.9.1. Chọn mẫu

Đa số các trường hợp nghiên cứu được tiến hành trên mẫu, tuy nhiên kết quả nghiên cứu phải có tính suy diễn cho quần thể nên mẫu nghiên cứu phải thực sự đại diện cho quần thể nghiên cứu. Chọn mẫu thiên lệch sẽ làm cho mẫu không đại diện, kết quả không đại diện được cho quần thể nghiên cứu mà chỉ thể hiện được ở mức độ mẫu nghiên cứu.

Loại bỏ hoàn toàn các nguồn tạo nên thiên lệch khi chọn mẫu là điều rất khó. Thiên lệch trong chọn mẫu điều tra có thể xảy ra khi người điều tra chọn mẫu ở các vùng dễ tiếp cận hay đối tượng dễ tiếp cận, dễ khai thác thông tin. Một trường hợp tạo nên thiên lệch khi chọn mẫu là người điều tra chọn mẫu thuận theo mong muốn chứng minh giả thuyết nghiên cứu theo ý muốn. Đây là nguyên nhân rất phổ biến trong nghiên cứu và cần được loại bỏ. Một trường hợp khác tạo nên sự thiên lệch khi chọn mẫu là nội dung nghiên cứu bị ảnh hưởng bởi màu sắc chính trị.

Dung lượng mẫu bao nhiêu luôn là câu hỏi thường trực đối với mọi nghiên cứu điều tra. Về mặt nguyên lý, với cùng một cách chọn mẫu thì độ chính xác của kết luận nghiên cứu sẽ tăng lên khi số mẫu nghiên cứu tăng lên. Lý tưởng nhất là chúng ta nghiên cứu trên quần thể. Tuy nhiên, đây là điều không tưởng (do hạn chế về nguồn lực) và chúng ta cũng không nhất thiết phải tiến hành nghiên cứu trên quần thể như vậy. Tốc độ cải thiện độ chính xác của kết luận nghiên cứu giảm khi dung lượng mẫu tăng, ví dụ tăng dung lượng mẫu từ 100 đến 250 thì độ chính xác tăng lên 3 lần, nhưng khi tăng dung lượng mẫu từ 250 lên 1.000 thì độ chính xác chỉ tăng gấp đôi.

Khi tính trạng nghiên cứu có sự biến động càng lớn, để đảm bảo độ chính xác của nghiên cứu thì cần có số lượng mẫu lớn và ngược lại. Dung lượng mẫu nghiên cứu cho các tính trạng định tính thường ít hơn các tính trạng định lượng. Các tính trạng có hệ số di truyền cao thì dung lượng mẫu nghiên cứu nhỏ hơn so với tính trạng có hệ số di truyền thấp. Dung lượng mẫu nghiên cứu cho các tính trạng sinh sản thường lớn hơn cho các tính trạng liên quan đến chất lượng sản phẩm.

Trong các nghiên cứu thiết kế thí nghiệm, xác định dung lượng mẫu nghiên cứu không phải là vấn đề khó khi chúng ta xác định được lực thống kê, mức sai khác có ý nghĩa... Tuy nhiên trong nghiên cứu điều tra, xác định dung lượng mẫu nghiên cứu không phải là vấn đề đơn giản do nghiên cứu điều tra thông thường có rất nhiều tính trạng nghiên cứu. Mỗi tính trạng nghiên cứu lại có một mức độ biến động khác nhau, do vậy cần số lượng mẫu có ý nghĩa khác nhau. Một số tác giả đề xuất dung lượng mẫu nghiên cứu dựa trên dung lượng quần thể (Bảng 9.1).

Bảng 9.1. Dung lượng mẫu nghiên cứu dựa theo dung lượng quần thể nghiên cứu

Dung lượng quần thể	Dung lượng mẫu
50	44
75	63
100	80
150	108
200	132
300	168
400	196
500	217
1000	277
3000	340
5000	356
10000	396

Một trong những công thức ước tính dung lượng mẫu nghiên cứu trong nghiên cứu điều tra là công thức của Slovin (1984):

$$n = N / (1 + Ne^2) \quad [9.1]$$

Trong đó:

- n: Số mẫu nghiên cứu
- N: Số lượng cá thể của quần thể
- e: sai số

Ví dụ khi $N = 18000$ và $e = 10\% = 0,1$

$$n = 18000 / (1 + 18000 \times 0,1^2)$$

$$n = 18000 / (1 + 18000 \times 0,01)$$

$$n = 18000 / (1 + 180)$$

$$n = 18000 / 181$$

$$n = 99,45$$

9.3.9.2. Cách chọn mẫu

Như đã đề cập trong chương 1, hai nguyên tắc cơ bản của nghiên cứu thí nghiệm và nghiên cứu điều tra đó là ngẫu nhiên và lặp lại. Nghiên cứu trên một dung lượng mẫu nhất định đảm bảo nguyên tắc lặp lại. Nguyên tắc ngẫu nhiên trong chọn mẫu nghiên cứu điều tra cũng cần được đảm bảo. Tất cả các mẫu nghiên cứu cần được chọn một cách ngẫu nhiên từ quần thể nghiên cứu. Một số phương pháp chọn mẫu ngẫu nhiên như sau:

- Mẫu ngẫu nhiên đơn

Sử dụng phương pháp chọn mẫu ngẫu nhiên đơn khi toàn bộ quần thể nghiên cứu tiềm năng đồng nhất. Phương pháp chọn mẫu ngẫu nhiên đơn đảm bảo nguyên tắc các đối tượng nghiên cứu tiềm năng có cơ hội/xác suất được lựa chọn giống nhau. Có nhiều cách khác nhau để thực hiện cách chọn mẫu này như sử dụng bảng số ngẫu nhiên, sử dụng hàm Random() hoặc Randombetween() trong một số phần mềm thống kê.

- Mẫu ngẫu nhiên phân tầng

Sử dụng phương pháp chọn mẫu ngẫu nhiên phân tầng thông thường được áp dụng khi quần thể nghiên cứu không đồng nhất. Mẫu nghiên cứu có thể chia ra thành từng nhóm phụ, ví dụ nhóm hộ nuôi bò Vàng, nhóm hộ nuôi bò lai F_1 (Vàng $\times$ Brahman), nhóm hộ nuôi bò Brahman thuần. Để có thể chọn mẫu nghiên cứu, chúng ta sử dụng kỹ thuật chọn mẫu ngẫu nhiên đơn, tuy nhiên thực hiện chọn mẫu ngẫu nhiên đơn trong trường hợp này chỉ được tiến hành trong mỗi nhóm hộ.

Sau khi đã chọn mẫu xong thì cần tiến hành ngay điều tra chính thức. Để cho việc điều tra đạt hiệu quả cao nhất cần tập huấn cho người điều tra, cần có kế hoạch điều tra cụ thể. Đặc biệt cần có biện pháp kiểm tra mức độ chính xác của thông tin thu thập được. Người điều tra cần kiểm tra các phiếu điều tra sau mỗi ngày điều tra.

9.3.10. Nhập số liệu

Sau khi đã thu thập số liệu bằng bản hỏi, chúng ta tiến hành nhập số liệu. Các số liệu điều tra thông thường được nhập và quản lý bởi Excel. Khi cần xử lý thì nạp tệp (file) dữ liệu vào các phần mềm thống kê để phân tích số liệu. Chúng ta cũng có thể sử dụng các phần mềm thống kê khác để nhập và quản lý dữ liệu như Statistics, SPSS, Access... Điều đặc biệt lưu ý là Excel thường bị giới hạn bởi số cột. Thông thường số lượng mẫu nghiên cứu không quá lớn để bị giới hạn bởi hàng. Khi số lượng biến nghiên cứu quá lớn thì chúng ta không nên sử dụng phần mềm Excel để nhập và quản lý số liệu.

Mã hóa biến nghiên cứu và giá trị của biến nghiên cứu phải được thực hiện trước khi nhập số liệu. Nguyên tắc của mã hóa phải đảm bảo dễ hiểu và thuận tiện cho việc thao tác và chia sẻ cơ sở dữ liệu. Đặc biệt khi phân tích số liệu bằng cú pháp và câu lệnh

thì mã hóa biến nghiên cứu và giá trị của biến càng đòi hỏi dễ hiểu và dễ nhớ. Mã hóa biến nghiên cứu và giá trị của biến thông thường được kết hợp với đặt nhãn (label) để có thể hiểu nghĩa của biến và giá trị của biến một cách đầy đủ.

Ví dụ 9.4. Câu hỏi: “Gia đình Ông/bà nuôi giống bò gì”

Câu trả lời:

- ☐ Bò Vàng
- ☐ Bò Lai
- ☐ Cả bò Vàng và bò Lai

Nếu như chúng ta nhập nguyên kết quả trả lời “Bò Vàng” hay “Bò Lai” hay “Cả bò Vàng và bò Lai” thì sẽ gây mất thời gian và dễ gặp sai sót. Để khắc phục hạn chế này chúng ta mã hóa. Một giải pháp mã hóa như sau: Thay vì nhập “Bò Vàng” chúng ta chỉ nhập “V”; thay vì nhập “Bò Lai” chúng ta chỉ nhập “L”; thay vì nhập “Cả bò Vàng và bò Lai” ta nhập “CH”. Chúng ta cũng có thể mã hóa ở dạng số. Thay vì nhập “Bò Vàng” chúng ta chỉ nhập số 1; thay vì nhập “Bò Lai” chúng ta chỉ nhập số 2; thay vì nhập “Cả bò Vàng và bò Lai” ta nhập số 3.

Để có thể nhập số liệu chúng ta cần xác định rõ:

- Đơn vị nghiên cứu, trong nghiên cứu thí nghiệm còn gọi là đơn vị thí nghiệm
- Số biến độc lập của nghiên cứu
- Số biến phụ thuộc của nghiên cứu

Để nhập số liệu đúng, đa số các phần mềm quản lý lý số liệu hiện tại tuân theo các nguyên tắc sau đây:

- Mỗi đơn vị nghiên cứu được nhập vào trong một hàng của worksheet. Tất cả các thông tin của mỗi đơn vị nghiên cứu đều phải được nằm cùng trong một hàng.
- Hàng thứ nhất của worksheet là hàng tên biến (độc lập và phụ thuộc).
- Mỗi biến độc lập hay biến phụ thuộc nằm trong một cột. Tất cả các thông tin về một biến phải được nằm trong một cột.

Sau khi hoàn tất việc nhập số liệu, chúng ta không nên vội vàng phân tích số liệu. Kinh nghiệm cho thấy nhập số liệu luôn có sai sót, do vậy các số liệu cần được kiểm tra sau khi được nhập. Để có thể kiểm tra sai sót khi nhập số liệu cần có sự liên kết giữa thông tin trên phiếu điều tra và thông tin đã được nhập trên máy tính. Mỗi phiếu điều tra cần có một mã phiếu. Mã phiếu này cũng cần được nhập vào máy tính để thuận tiện cho việc tra cứu.

Có rất nhiều trình ứng dụng khác nhau để kiểm tra độ chính xác của việc nhập số liệu, như dùng RANGE trong phần mềm Excel, Box Plot trong SPSS, hay thống kê mô tả trong các phần mềm thống kê.

Bảng 9.2. Trích đoạn nhập số liệu trong đề tài “Ảnh hưởng của giống và vùng sinh thái đến khả năng sinh sản của bò Vàng và bò lai giữa bò Vàng và bò lai Brahman nuôi tại Bình Định”

Mã số	Xã	Huyện	Vùng sinh thái	Giống bò	Tỷ lệ máu bò lai (%)	Tuổi bò (tháng)	Số lứa đã đẻ	Tuổi động dục lần đầu (tháng)	Tuổi phối lần đầu (tháng)
A1	nm	an	db	I	50	54	1	36	36
A2	nm	an	db	I	50	60	2	30	30
A3	nm	an	db	I	75	78	3	36	36
A4	nm	an	db	I	50	69	3	26	26
A5	nm	an	db	I	75	59	2	34	34
A6	nm	an	db	I	75	53	1	30	30
A7	nm	an	db	I	75	76	3	28	28
A8	nm	an	db	I	75	64	2	32	32
A9	nm	an	db	I	75	75	3	34	34
A10	nm	an	db	I	50	92	4	32	32
A11	nm	an	db	I	75	66	2	36	36
A12	nm	an	db	I	75	62	2	32	32
A13	nm	an	db	I	50	66	3	28	28
A14	nm	an	db	I	50	65	3	26	26
A15	nm	an	db	I	50	58	2	30	30
A16	nm	an	db	I	50	100	4	32	32
A17	nm	an	db	I	50	84	4	26	26

Nguồn: Lê Đình Phùng (2009). Ảnh hưởng của giống và vùng sinh thái đến khả năng và hiệu quả chăn nuôi bò tại tỉnh Bình Định. *Tạp chí Nông nghiệp và Phát triển Nông thôn* 8: 67-71.

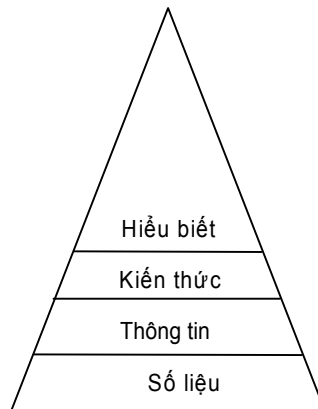
Trong nghiên cứu điều tra, nhiều khi chúng ta gặp các giá trị ngoại lệ. Việc đầu tiên cần kiểm tra là giá trị ngoại lệ đó có tồn tại không bằng cách kiểm tra lại giá trị ngoại lệ đó trong bản hỏi. Nếu thực sự tồn tại thì không nên vội vàng loại bỏ giá trị ngoại lệ. Giá trị ngoại lệ thông thường là cơ sở cho các nghiên cứu giải thích sự tồn tại của giá trị ngoại lệ. Trong nghiên cứu thí nghiệm liên quan đến khả năng sản xuất của vật nuôi, các giá trị ngoại lệ thông thường là cơ sở cho nghiên cứu chọn lọc cá thể.

9.3.11. Phân tích số liệu

Phân tích số liệu có ý nghĩa rất quan trọng trong nghiên cứu. Sơ đồ 9.5 mô phỏng ý nghĩa của phân tích số liệu. Nhờ có phân tích số liệu mà chúng ta có thông tin, có kiến thức và có được sự hiểu biết. Đặc biệt đối với nghiên cứu điều tra bằng bản hỏi thì phân tích số liệu chính là bản anh hùng ca của nghiên cứu. Trong thực tế, thông thường các nhà nghiên cứu chưa phân tích một cách thấu đáo các số liệu nghiên cứu điều tra.

Có sự liên quan chặt chẽ giữa kỹ năng phân tích số liệu và kỹ năng soạn bản hỏi. Bản hỏi sẽ ngắn, trực tiếp vào vấn đề cần hỏi nếu nhóm nghiên cứu có kỹ năng phân tích số liệu tốt. Kỹ năng phân tích số liệu tốt cũng có nghĩa là bản hỏi sẽ tập trung vào những thông tin trọng tâm hơn là tập trung vào những thông tin thú vị.

Để phân tích số liệu đúng chúng ta cần phân biệt rõ ràng giữa biến độc lập và biến phụ thuộc. Trong nghiên cứu thí nghiệm phân biệt giữa biến độc lập và biến phụ thuộc không khó khăn như trong nghiên cứu điều tra. Trong nghiên cứu điều tra, một biến vừa có thể là biến độc lập và cũng có thể là biến phụ thuộc.



Biểu đồ 9.5. Mô hình mô phỏng ý nghĩa của phân tích số liệu

Trong nghiên cứu điều tra, trước khi phân tích số liệu chúng ta có thể cần phải tính toán giá trị của một số biến, đặc biệt là các biến tổng hợp. Thông thường người được phỏng vấn không thể cung cấp một cách chính xác giá trị của các biến tổng hợp. Để đảm bảo độ chính xác thông thường chúng ta thu thập các thông tin đơn lẻ cung cấp cấp thông tin đầu vào cho việc tính giá trị của biến tổng hợp. Ví dụ để tính thu nhập từ chăn nuôi, chúng ta trước hết cần tính toán thu nhập từ chăn nuôi lợn, chăn nuôi gia cầm... Để tính thu nhập từ chăn nuôi lợn chúng ta cần tính các giá trị đầu ra và đầu vào từ chăn nuôi lợn...

Như đã đề cập trong các chương trước, chúng ta có thể phân tích thống kê mô tả với các tham số:

- Đo lường mức độ tập trung
 - Trung bình thực (Phân bố chuẩn)
 - Trung bình nhân (Phân bố lệch)
 - Trung vị
- Đo lường mức độ phân tán
 - Độ lệch chuẩn
 - Phương sai
 - Hệ số biến dị
 - Phạm vi biến động
- Đo lường mối quan hệ
 - $r = \text{cov}(x,y)/(\text{sd}(x) \times \text{sd}(y))$

Chúng ta có thể sử dụng các mô hình thống kê để phân tích các số liệu điều tra.

$$Y = f(x) + g(e) \quad [9.2]$$

Trong đó Y là biến phụ thuộc, hoặc biến cần được giải thích. Y có thể là biến liên tục và có thể là biến phân tử. Khi Y là biến phân tử (Có hoặc Không, 1 hoặc 0, ...) chúng ta sử dụng Logit hoặc Multinomial logit model để phân tích ảnh hưởng của biến độc lập đến biến phụ thuộc.

$f(x)$ là hàm tuyến tính hoặc phi tuyến tính. $f(x)$ mô tả mối quan hệ giữa y và x. $g(e)$ là một hàm ngẫu nhiên. Biến x có thể là biến liên tục khi đó phân tích công thức 9.2. chính là phân tích hồi quy (regression). Khi x là biến phân nhóm khi đó phân tích công thức 9.2 chính là phân tích phương sai (ANOVA).

- Mô hình hồi quy tuyến tính: $f(x)$ có thể hồi quy tuyến tính với một hoặc nhiều hơn một biến x

$$f(x) = a + bx \quad [9.3]$$

$$f(x) = a + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad [9.4]$$

- Mô hình phân tích phương sai $f(x)$ được quyết định bởi các mức của nhân tố. Ví dụ trong trường hợp phân tích 2 nhân tố, mô hình phân tích được thể hiện [9.5].

$$f(x) = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \quad [9.5]$$

Sau khi đã phân tích số liệu thì chúng ta có thể báo cáo kết quả nghiên cứu Các hình thức báo cáo nghiên cứu có thể là luận văn, bài báo khoa học, bài trình bày tại hội thảo. Trong phạm vi của giáo trình này, nội dung viết khoa học sẽ không được trình bày.

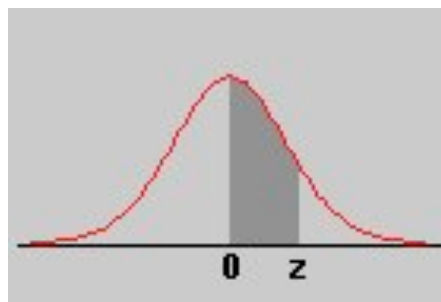
9.4. Bài tập

1. Xây dựng bản hỏi để thu thập thông tin đánh giá tình hình chăn nuôi lợn nái ngoại tại huyện Phong Điền tỉnh Thừa Thiên Huế.
2. Xây dựng bản hỏi để thu thập thông tin đánh giá ảnh hưởng của việc trồng cỏ nuôi bò đến tình hình kinh tế, xã hội và môi trường tại Thừa Thiên Huế.
3. Xây dựng bản hỏi để đánh giá các yếu tố ảnh hưởng đến chấp nhận tiến bộ kỹ thuật trong chăn nuôi của người dân.
4. Xây dựng bản hỏi để đánh giá ảnh hưởng của chương trình cải tạo đàn bò đến khả năng và hiệu quả chăn nuôi bò ở Thừa Thiên Huế.
5. Xây dựng bản hỏi để đánh giá một số nhân tố ảnh hưởng đến khả năng sinh sản của lợn nái ngoại trong các trang trại tập trung của tỉnh Quảng Bình.
6. Xây dựng các bản hỏi để đánh giá các yếu tố ảnh hưởng đến sự bùng phát và lan truyền dịch cúm gia cầm ở Thừa Thiên Huế.
7. Tại sao ước tính số mẫu điều tra trong nghiên cứu điều tra khó hơn so với ước tính số đơn vị thí nghiệm trong nghiên cứu thí nghiệm?

Phụ lục 1. Bảng phân bố chuẩn hóa (Z)

Các giá trị trong bảng dưới đây thể hiện diện tích phía dưới đường phân bố chuẩn hóa (standard normal curve) cho các giá trị giữa 0 và giá trị z . Ví dụ để xác định diện tích phía dưới đường phân bố chuẩn hóa giữa 0 và giá trị z là 2,31, chúng ta tra bảng với giá trị ở hàng = 2,3 và giá trị ở cột là 0,01. Kết quả diện tích phía dưới đường phân bố là 0,4896.

Diện tích giữa 0 và z



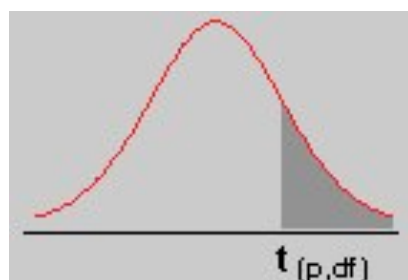
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Phụ lục 2. Bảng phân bố Student's t

Các giá trị trong hàng đầu tiên của bảng t là diện tích đuôi bên phải của sự phân bố đối với giá trị t trong bảng, các giá trị trong cột đầu tiên là độ tự do. Ví dụ để xác định giá trị quyết định từ phân bố t ở mức $\alpha = 0,05$ với 8 độ tự do ta có $t_{(0,05, 8)} = 1,859548$.

Bảng t với kiểm tra một đuôi bên phải



df\p	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0005
1	0,324920	1,000000	3,077684	6,313752	12,70620	31,82052	63,65674	636,6192
2	0,288675	0,816497	1,885618	2,919986	4,30265	6,96456	9,92484	31,5991
3	0,276671	0,764892	1,637744	2,353363	3,18245	4,54070	5,84091	12,9240
4	0,270722	0,740697	1,533206	2,131847	2,77645	3,74695	4,60409	8,6103
5	0,267181	0,726687	1,475884	2,015048	2,57058	3,36493	4,03214	6,8688
6	0,264835	0,717558	1,439756	1,943180	2,44691	3,14267	3,70743	5,9588
7	0,263167	0,711142	1,414924	1,894579	2,36462	2,99795	3,49948	5,4079
8	0,261921	0,706387	1,396815	1,859548	2,30600	2,89646	3,35539	5,0413
9	0,260955	0,702722	1,383029	1,833113	2,26216	2,82144	3,24984	4,7809
10	0,260185	0,699812	1,372184	1,812461	2,22814	2,76377	3,16927	4,5869
11	0,259556	0,697445	1,363430	1,795885	2,20099	2,71808	3,10581	4,4370
12	0,259033	0,695483	1,356217	1,782288	2,17881	2,68100	3,05454	4,3178
13	0,258591	0,693829	1,350171	1,770933	2,16037	2,65031	3,01228	4,2208
14	0,258213	0,692417	1,345030	1,761310	2,14479	2,62449	2,97684	4,1405

df/p	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0005
15	0,257885	0,691197	1,340606	1,753050	2,13145	2,60248	2,94671	4,0728
16	0,257599	0,690132	1,336757	1,745884	2,11991	2,58349	2,92078	4,0150
17	0,257347	0,689195	1,333379	1,739607	2,10982	2,56693	2,89823	3,9651
18	0,257123	0,688364	1,330391	1,734064	2,10092	2,55238	2,87844	3,9216
19	0,256923	0,687621	1,327728	1,729133	2,09302	2,53948	2,86093	3,8834
20	0,256743	0,686954	1,325341	1,724718	2,08596	2,52798	2,84534	3,8495
21	0,256580	0,686352	1,323188	1,720743	2,07961	2,51765	2,83136	3,8193
22	0,256432	0,685805	1,321237	1,717144	2,07387	2,50832	2,81876	3,7921
23	0,256297	0,685306	1,319460	1,713872	2,06866	2,49987	2,80734	3,7676
24	0,256173	0,684850	1,317836	1,710882	2,06390	2,49216	2,79694	3,7454
25	0,256060	0,684430	1,316345	1,708141	2,05954	2,48511	2,78744	3,7251
26	0,255955	0,684043	1,314972	1,705618	2,05553	2,47863	2,77871	3,7066
27	0,255858	0,683685	1,313703	1,703288	2,05183	2,47266	2,77068	3,6896
28	0,255768	0,683353	1,312527	1,701131	2,04841	2,46714	2,76326	3,6739
29	0,255684	0,683044	1,311434	1,699127	2,04523	2,46202	2,75639	3,6594
30	0,255605	0,682756	1,310415	1,697261	2,04227	2,45726	2,75000	3,6460
∞	0,253347	0,674490	1,281552	1,644854	1,95996	2,32635	2,57583	3,2905