BÀI TOÁN SO SÁNH MỞ RỘNG

§ 1. SO SÁNH NHIỀU TỶ LỆ

Trong chương trước chúng ta đã xét bài toán so sánh tỷ lệ cá thể có đặc tính A trong hai tập hợp chính. bấy giờ chúng ta sẽ mở rộng bài toán này bằng cách xét bài toán so sánh đồng thời tỷ lệ cá thể có đặc tính A giữa nhiều tập hợp chính.

Giả sử ta có k tập hợp chính $H_1, H_2,... H_k$. Mỗi cá thể của chúng có thể mang hay không mang đặc tính A.

Gọi p_1 là tỷ lệ có thể mang đặc tính A trong tập hợp chính H_i (i = 1, 2, ...k).

Các tỷ lệ này được gọi là các tỷ lệ lý thuyết mà chúng ta chưa biết.

Ta muốn kiểm định giả thiết sau:

H_o: $p_1 = p_2 = ... = p_k$ (tất cả các tỷ lệ này bằng nhau).

Từ mỗi tập hợp chính H_i ta rút ra một ngẫu nhiên có kích thước n_i, trong đó chúng ta thấy có m_i cá thể mang đặc tính A. các dữ liệu này được trình bày trong bảng sau đây:

Mẫu	**1**	**2**	**...**	**k**	**Tổng**
Có A	m_1	m_2	...	m_k	m
Không A	l_1	l_2	...	l_k	l
Tổng	n_1	n_2	...	n_k	$N = m + l = \sum n_i$

Nếu giả thiết

$$H_o: p_1 = p_2 = ... = p_k = p$$

Là đúng thì tỷ lệ chung p được ước lượng bằng tỷ số giữa số cá thể đặc tính A của toàn bộ k mẫu gộp lại trên tổng số cá thể của k mẫu gộp lại.

$$p = \frac{m}{N}$$

Tỷ lệ cá thể không có đặc tính A được ước lượng bởi

$$q = 1 - p = \frac{l}{N}$$

Khi đó số cá thể có đặc tính A trong mẫu thứ i (mẫu rút từ tập hợp chính H_i) sẽ xấp xỉ bằng

$$m_i = n_i p = \frac{n_i m}{N}$$

và số cá thể không có đặc tính A trong mẫu thứ i sẽ xấp xỉ bằng

$$\hat{l}_i = n_i q = n_i \frac{l}{N}$$

Các số m_i và $\hat{l}_i$ được gọi là các tần số lý thuyết (TSLT), còn các số m_i, l_i được gọi là các tần số quan sát (TSQS).

Ta quyết định bác bỏ H_o khi TSLT cách xa TSQS một cách "bất thường". Khoảng cách giữa TSQS và TSLT được đo bằng test thống kê sau đây:

$$T = \sum_{i=1}^{k} \frac{\left(m_i - m_i\right)^2}{m_i} + \sum_{i=1}^{k} \frac{\left(l_i - \hat{l}_i\right)^2}{\hat{l}_i}$$

Người ta chứng minh được rằng nếu H_o đúng và các tần số lý thuyết không nhỏ thua 5 thì T sẽ có phân bố xấp xỉ phân bố χ^2 với k – 1 bậc tự do. Thành thử miền bác bỏ H_o có dạng $\{T > c\}$, ở đó c được tìm từ điều kiện $P\{T > c\} = \alpha$. Vậy c chính là phân vị mức α của phân bố χ^2 với k – 1 bậc tự do.

Chú ý. Test thống kê T có thể biến đổi như sau.

Ta có:

$$\left(l_i - \hat{l}_i\right)^2 = \left[n_i - m_i - n_i\left(1 - p\right)\right]^2 = \left(m_i - n_i p\right)^2 = \left(m_i - m_i\right)^2$$

Do đó

$$T = \sum \left(m_i - m_i \right)^2 \left(\frac{1}{m_1} + \frac{1}{\hat{l}_i} \right)$$

$$= \sum \left(m_i - m_i \right)^2 \left(\frac{1}{n_i p_1} + \frac{1}{n_i q} \right)$$

$$= \sum_{i=1}^{k} \frac{\left(m_i - m_i \right)^2}{n_i pq} = \sum \frac{m_i^2}{n_i pq} - 2\sum \frac{m_i m_i}{n_i pq} + \sum \frac{m_o^2}{n_i pq}$$

Chú ý rằng

$$\sum \frac{m_i m_i}{n_i pq} = \frac{1}{q} \sum m_i = \frac{m}{q}; \quad \sum \frac{m_1^2}{n_i pq} = \frac{1}{q} \sum m_i = \frac{m}{q}$$

Vậy

$$T = \frac{1}{pq} \sum \frac{m_i^2}{n_i} - \frac{m}{q} = \frac{1}{pq} \sum \frac{m_i^2}{n_i} - N\frac{p}{q} = \frac{N^2}{ml} \sum \frac{m_i^2}{n_i} - N\frac{m}{l}$$

Nếu sử dụng công thức này ta sẽ không cần tính các tần số lý thuyết, do đó nó được dùng trong thực hành.

Ví dụ 1. So sánh tác dụng của 6 mẫu thuốc thử nghiệm trên 6 lô chuột, kết quả thu được như sau:

Mẫu thuốc	**1**	**2**	**3**	**4**	**5**	**6**	**Tổng**
Số sống	79	82	77	83	76	81	478
Số chết	21	18	23	17	24	19	122
Tổng	100	100	100	100	100	100	600

Ta muốn kiểm định giả thiết

H_o: Tỷ lệ chết trong 6 mẫu thuốc là như nhau

Đối thiết H_1: Tỷ lệ chết trong 6 mẫu thuốc là khác nhau

Giải

Ta có $$T = \frac{600^2}{(478)(122)} \left[\frac{79^2}{100} + \frac{82^2}{100} + \cdots + \frac{81^2}{100} \right] - \frac{(600)(478)}{122}$$

$$= 2353,24 - 2350,81 = 2,42$$

Với mức ý nghĩa $\alpha = 5\%$, tra bảng phân bố χ^2 với 5 bậc tự do ta có

$$\chi^2_{0,05} = 11,07$$

Vì T < c nên ta chấp nhận H_o. J

Ví dụ 2. Có 4 thầy giáo A, B, C, D cùng dạy một giáo trình thống kê. Ban chủ nhiệm khoa muốn tìm hiểu chất lượng dạy của 4 thầy này nên đã làm một cuộc khảo sát. Kết quả như sau:

Thầy / Kết quả	A	B	C	D	Tổng
Đạt	60	75	150	125	410
Không đạt	40	75	50	75	240
Tổng	100	150	200	200	650

Với mức ý nghĩa $\alpha = 0,01$ có thể cho rằng tỷ lệ học sinh đỗ trong các học sinh đã học các thầy trên là như nhau hay không?

Giải. Ta có

$$T = \frac{(650)^2}{(410)(240)}\left[\frac{60^2}{100} + \frac{75^2}{150} + \frac{150^2}{200} + \frac{125^2}{200}\right] - \frac{(650)(410)}{240}$$
$$= 1134,07 - 1110,41 = 23,65$$

Số bậc tự do là 3 và $\chi^2_{0,01} = 11,343$. Vì T > c nên ta bác bỏ giả thuyết H_o. Tỷ lệ học sinh đỗ của các thầy A, B, C, D như nhau.

§ 2. SO SÁNH CÁC PHÂN SỐ

Xét một bộ A gồm r tính trạng, $A = (A_1, A_2, ...A_r)$, trong đó mỗi cá thể của tập hợp chính H có và chỉ có một trong các tính trạng (hay phạm trù) A_i.

Gọi p_i (i = 1, 2, ... r) là tỷ lệ cá thể tính trạng A_i trong tập hợp chính H. Khi đó véctơ $\pi = (p_1, p_2, ...p_r)$ được gọi là phân bố của A trong tập hợp chính H.

Chẳng hạn, mọi người đi làm có thể sử dụng một trong các phương tiện sau: đi bộ, đi xe đạp, đi xe máy, đi xe buýt. Trong thành phố X có 18% đi bộ, 32% đi xe đạp, 40% đi xe máy và 10% đi xe buýt. Như vậy π = (0,18; 0,32; 0,4; 0,1) là phân bố của cách đi làm (A) trong tập hợp các dân cư của thành phố X.

Tương tự mỗi người có thể được xếp vào 1 trong 3 phạm trù sau: rất hạnh phúc, bất hạnh, hoặc có thể được xếp vào 1 trong 3 lớp sau: dưới 25

tuổi, trong khoảng từ 25 đến 45 tuổi, trên 45 tuổi... có thể dẫn ra rất nhiều ví dụ tương tự như vậy.

Giả sử $(p_1, p_2,...p_r)$ là phân bố của $(A_1, A_2,...A_r)$ trong tập hợp chính H và $(q_1, q_2,...q_r)$ là phân bố của $A = (A_1, A_2,...A_r)$ trong tập hợp chính Y. Ta nói $(A_1, A_2...A_r)$ có phân bố như nhau trong X và Y nếu $(p_1, p_2,...p_r) = (q_1, q_2,...r_r) \Leftrightarrow p_1 = q_1,...p_r = q_r$.

Chúng ta muốn kiểm định xem $A = (A_1, A_2,...A_r)$ có cùng phân số trong X và Y hay không dựa trên các mẫu ngẫu nhiên rút từ X và Y.

Tổng quát hơn, giả sử ta có k tập hợp chính $H_1, H_2,...H_k$. Gọi $\pi^i = \left(p_1^i, p_2^i, \ldots p_r^i\right)$ là phân bố của $A = (A_1, A_2,...A_r)$ trong tập hợp chính H_i.

Ta muốn kiểm định giả thuyết sau

$H_o: \pi^1 = \pi^2 = \ldots = \pi^k$ (Các phân bố này là như nhau trên các tập hợp chính H_i).

Chú ý rằng H_o tương đương với hệ đẳng thức sau:

$$\begin{cases} p_1^1 = p_1^2 = \ldots = p_1^k \\ p_2^1 = p_2^2 = \ldots = p_2^k \\ p_i^1 = p_i^2 = \ldots = p_i^k \\ p_r^1 = p_r^2 = \ldots = p_r^k \end{cases}$$

Từ mỗi tập hợp chính chúng ta chọn ra một mẫu ngẫu nhiên. Mẫu ngẫu nhiên chọn từ tập hợp chính H_i được gọi là mẫu ngẫu nhiên thứ i (i = 1, 2,... k).

Giả sử trong mẫu ngẫu nhiên thứ i

Có n_{1i} cá thể có tính trạng A_1

n_{2i} cá thể có tính trạng A_2

.....

n_{ri} cá thể có tính trạng A_r

Ta xắp xếp cá số liệu đó thành bảng sau đây.

Mẫu / Tính trạng	1	2		J		K	Tổng số
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_{10}
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_{20}

...	...	...	...	...	...	...	...
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_{i0}
...	...	...	...	...	...	...	...
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rk}	n_{r0}
Tổng số	n_{o1}	n_{o2}	...	n_{oj}	...	n_{ok}	n

Ký hiệu $n_{io} = \sum_{j=1}^{k} n_{ij}$

$$n_{oj} = \sum_{i=1}^{r} n_{ij}$$

Như vậy n_{oj} là kích thước của mẫu thứ j, còn n_{io} là tổng số cá thể có tính trạng A_i trong toàn bộ k mẫu đang xét

$$n = \sum_{i=1}^{r} n_{io} = \sum_{j=1}^{k} n_{oj}$$

Là tổng số tất cả các cá thể của k mẫu đang xét.

Nếu giả thiết H_o là đúng nghĩa là

$$\begin{cases} p_1^1 = p_1^2 = \ldots = p_1^k = p_1 \\ p_2^1 = p_2^2 = \ldots = p_2^k = p_2 \\ \text{---------- } \\ p_i^1 = p_i^2 = \ldots = p_i^k = p_i \\ \text{---------- } \\ p_r^1 = p_r^2 = \ldots = p_r^k = p_r \end{cases}$$

thì các tỷ lệ chung $p_1, p_2, ...p_r$ được ước lượng bởi:

$$p_i = \frac{n_{io}}{n}$$

Đó ước lượng cho xác suất để một cá thể có mang tính trạng A_i. khi đó số cá thể có tính trạng A_i trong mẫu thứ j sẽ xấp xỉ bằng

$$n_{ij} = n_{oj} p_i = \frac{n_{oj} n_{io}}{n}$$

Các số $n_{ij} (i = 1,2,...r;\ j = 1,2,...k)$

được gọi là các tần số lý thuyết (TSLT), các số n_{ij} được gọi là các tần số quan sát (TSQS).

Ta quyết định bác bỏ H_o khi các TSLT cách xa TSQS một cách bất thường. Khoảng cách giữa TSQS và TSLT được đo bằng test thống kê sau đây

$$T = \sum_{f=1}^{k}\sum_{i=1}^{r}\frac{\left(n_{ij} - n_{ij}\right)^2}{n_{ij}} = \sum\frac{(TSQS - TSLT)^2}{TSLT}$$

Người ta chứng minh được rằng nếu H_o đúng và các TSLT không nhỏ hơn 5 thì T sẽ có phân bố xấp xỉ phân bố χ^2 với (k-1)(r-1) bậc tự do. Thành thử miền bác bỏ có dạng {T > c} ở đó c được tìm từ điều kiện P{T > c} = α. Vậy c là phân vị mức α của phân bố χ^2 với (k-1)(r-1) bậc tự do.

Chú ý. T có thể biến đổi thành các dạng sau đây.

Ta có $$\frac{\left(n_{ij} - n_{ij}\right)^2}{n_{ij}} = \frac{n_{ij}^2}{n_{ij}} - 2n_{ij} + n_{ij}$$

Để ý rằng: $\sum\sum n_{ij} = \sum\sum n_{ij} = n$

Vậy $$T = \sum\frac{n_{ij}^2}{n_{ij}} - 2n + n = \sum\frac{n_{ij}^2}{n_{ij}} = n\sum\frac{n_{ij}^2}{n_{io}n_{oj}} - n = n\left\{\sum\frac{n_{ij}^2}{n_{io}n_{oj}} - 1\right\} \qquad (1)$$

Với công thức này ta không phải tính các TSLT n_{ij}, do đó thường được sử dụng trong thực hành.

Ví dụ 3. Người ta muốn so sánh số băng trên vỏ của ba loài ốc sên rừng I, II và III. Số liệu nghiên cứu được cho ở bảng sau:

Số băng trên vỏ \ Loài	I	II	III	Tổng số
0	49	31	126	206
1 hoặc 2	33	20	56	109
3 hoặc 4	52	20	83	155
5 trở lên	35	29	109	173
Tổng số	169	100	374	643

Hỏi có thể cho rằng số băng trên vỏ có phân phối như nhau trên cả ba loài ốc sên này không? Chọn mức ý nghĩa là 5%.

Giải. Ta tính thống kê T theo công thức (1)

$$T = 643\left[\frac{49^2}{(169)(206)} + \frac{31^2}{(100)(206)} + \frac{126^2}{(374)(206)} + \right.$$

$$+ \frac{33^2}{(169)(109)} + \frac{20^2}{(109)(100)} + \frac{56^2}{(109)(374)} +$$

$$\left. + \cdots + \frac{29^2}{(100)(173)} + \frac{109^2}{(374)(173)} - 1\right] \approx 10,4$$

Tra bảng phân bố χ^2 với bậc tự do (3 – 1)(4 – 1) = 6, ta tìm được

$$c = \chi^2_{0,05} = 12,592$$

Giá trị này lớn hơn T. vậy chúng ta chấp nhận H_o: Số băng trên vỏ có phân bố như nhau đối với cả 3 loài ốc sên rừng.

Ví dụ 4. đài truyền hình việt nam muốn thăng dò ý kiến khán giả về thời lượng phát sóng phim truyện Việt Nam hàng tuần. Phiếu thăm dó đặt ra 4 mức.

A_1: Tăng thời lượng phát sóng

A_2: Giữ như cũ

A_3: Giảm

A_4: Không ý kiến

Đài đã tiến hành thăm dò ba nhóm xã hội khác nhau: công nhân, nông dân, trí thức. Kết quả cuộc thăm dò như sau:

Tầng lớp / Ý kiến	Công nhân	Nông dân	Trí thức	Tổng số
Tăng	100	300	20	420
Như cũ	200	400	30	630
Giảm	50	80	5	135
Không ý kiến	30	70	5	105
Tổng số	380	850	60	1290

Với mức ý nghĩa $\alpha = 5\%$, có sự khác nhau về ý kiến trong các tầng

lớp xã hội trên hay không?

Giải. Tần số lý thuyết của ô "trí thức không ý kiến" là $\frac{(60)(105)}{1290} = 4,88$, bé hơn 5 do đó điều kiện cho phép áp dụng tiêu chuẩn "khi bình phương" không được thoả mãn. Để khắc phục khó khăn này có hai cách. Hoặc là ghép dòng cuối cùng với một dòng nào đó, hoặc là ghép cột cuối cùng với một cột nào đó.

Tuy nhiên rất khó ghép dòng cuối cùng "không ý kiến" với một dòng nào đó cho hợp lý. "Không ý kiến" khác rất nhiều với việc "có bày tỏ ý kiến của mình". Hợp lý hơn ta ghép cột cuối cùng "trí thức" với cột "công nhân" vì trí thức có vẽ gần với công nhân hơn là nông dân (đều ở khu vực thành thị). Như vậy ta có bảng mới sau:

Tầng lớp / Ý kiến	Công nhân Và trí thức	Nông dân	Tổng số
Tăng	120	300	420
Như cũ	230	400	630
Giảm	55	80	135
Không ý kiến	35	70	105
Tổng số	440	850	1290

Sử dụng công thức tìm được

$$T = 1290\left[\frac{120^2}{(440)(220)} + \cdots + \frac{70^2}{(850)(105)} - 1\right] \approx 10,059$$

Tra bảng phân bố χ^2 ở mức 5% với bậc tự do là (2 – 1)(4 – 1) = 3, ta tìm được

$$\chi^2_{0,05} = 7,815$$

Số này bé hơn T. vậy ta kết luận rằng về thời lượng phát sóng phim Việt Nam có một sự khác nhau về ý kiến giữa hai tầng lớp xã hội: nông dân và công nhân viên chức.

Chú thích sử dụng Minitab

Để sử dụng Minitab thực hiện tiêu chuẩn χ^2 ta cần làm như sau. Các tần số quan sát được nhập vào dưới dạng các cột số liệu, chẳng hạn các

cột C_1, C_2, C_3 và C_4 bằng lệnh READ. Sau đó chúng ta đánh lệnh

CHIQUARE C1 – C4

Minitab sẽ cho ta trên màn hình các TSQS, TSLT, giá trị của test thống kê "Khi bình phương" T và số bậc tự do. Ta chỉ cần tra bảng phân bố χ^2 để tìm hằng số c và so sánh nó với giá trị của T.

Sau đây là ví dụ về một bảng mà Minitab cho ta trên màn hình:

```
MTB >  READ C1 – C4
       3 ROWS READ
MTB >  END
MTB >
MTB >  CHISQUARE C1 – C4
```

	C1	C2	C3	C4	Total
1	34	47	63	68	182
	36.79	42.64	66.42	36.14	
2	26	36	57	42	161
	32.55	37.73	58.75	31.97	
3	53	48	84	31	216
	43.66	50.62	78.83	42.89	
Total	113	131	204	111	559

```
Chisq = 11.299
DF = 6
MTB >
```

§ 2. PHÂN TÍCH PHƯƠNG SAI MỘT NHÂN TỐ

Trong chương 5 chúng ta xét bài toán so sánh giá trị trung bình của hai tập hợp chính. Trong mục này chúng ta xét bài toán tổng quát; so sánh đồng thời các giá trị trung bình của nhiều tập hợp chính.

Giả sử ta có k ĐLNN có phân bố chuẩn X_1, X_2, ... X_k, trong đó $X_i \sim N(\mu_i, \sigma_i^2)$.

Các giá trị trung bình μ_i và phương sai σ_i^2 đều chưa biết. Tuy nhiên chúng ta giả thiết rằng các phương sai bằng nhau:

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

Chúng ta muốn kiểm định xem liệu các giá trị trung bình μ_i này có

như nhau hay không:

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

Trong thốn gkê vấn đề trên thường được xem xét dưới góc độ sau đây.

Giả sử chúng ta quan tân đến một nhân tố X (factor) nào đó. Nhân tố X có thể xem xét ở k mức khác nhau. Ký hiệu X_i là hiệu quả của việc tác động nhân tố X ở mức i đối với cá thể. Như vậy μ_i là hiệu quả trung bình của nhân tố X ở mức i. chúng ta muốn biết khi cho nhân tố X thay đổi các mức khác nhau thì điều đó có ảnh hưởng hay không tới hiệu quả trung bình.

Ví dụ.

a) Chúng ta muốn nghiên cứu ảnh hưởng của giống tới năng suất cây trồng. Nhân tố đây là giống. Các loại giống khác nhau là các nức của nhân tố. Hiệu quả của giống lên năng suất cây trồng được đo bằng sản lượng của cây trồng. Như vậy X_i chính là sản lượng của giống i và μ_i là sản lượng trung bình của giống i.

b) Giả sử rằng có 4 giáo sư Toán A, B, C, D đang dạy một giáo trình xác suất cho năm thứ nhất. Nhà trường muốn tìm hiểu xem điểm thi trung bình của các sinh viên thụ giáo các giáo sư này có khác nhau hay không. Trong bối cảnh này, nhân tố là giáo sư. Mỗi giáo sư cụ thể là một mức của nhân tố. Hiệu quả của giáo sư A đối với cá thể (sinh viên) được đo bằng điểm thi của sinh viên đó. Như vậy X_A là điểm thi trung bình của tất cả các sinh viên này. Nhà trường muốn kiểm định giả thiết.

$$\mu_A = \mu_B = \mu_C = \mu_D$$

Giả sử $\{x_1, x_2, ...x_{n_1 1}\}$ là một mẫu có kích thước n_1 rút ra từ tập hợp chính các giá trị của X_1; $\{x_{12}, x_{22}, ...x_{n_2 2}\}$ là một mẫu kích thước rút ra từ tập hợp chính các giá trị của X_2,..., $\{x_{1k}, x_{2k}, ...x_{n_k k}\}$ là một mẫu kích thước n_k rút ra từ tập hợp chính các giá trị của X_k. các số liệu thu được trình bày thành bảng ở dạng sau đây:

Các mức nhân tố				
1	2	...	k	$n = \sum_{i=1}^{k} n_i$
x_{11}	x_{12}	...	n_{1k}	
x_{21}	x_{22}	...	n_{2k}	

	...	...	...	...	
	$x_{n_1 1}$	$x_{n_2 2}$	...	$x_{n_k k}$	
Tổng số	T_1	T_2	...	T_k	$T = \sum_{i=1}^{k} T_k$
Trung bình	$\bar{x}_1$	$\bar{x}_2$	...		$\bar{x} = \frac{T}{n}$

Ta đưa ra một số kí hiệu sau

*) Trung bình của mẫu thứ i (tức là mẫu ở cột thứ i trong bảng trên):

$$\bar{x}_i = \frac{T_i}{n_i} = \frac{\sum_{j=1}^{n_i} x_{ji}}{n_i}$$

*) Trung bình chung

$$\bar{x} = \frac{T}{n} = \frac{\sum\sum x_{ij}}{n} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_j} x_{ij}}{n}$$

ở đó $n = n_1 + n_2 + ... + n_k$;

$T = T_1 + T_2 + ... + T_k$.

*) Tổng bình phương chung ký hiệu là SST (viết tắt là chữ Total Sum of Squares) được tính theo công thức sau:

$$STT = \sum_{i=1}^{n_1}(x_{i1} - \bar{x})^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x})^2 + \cdots + \sum_{i=1}^{n_k}(x_{ik} - \bar{x})^2$$
$$= \sum_{j=1}^{n_k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$$

có thể chứng minh rằng

$$STT = \sum_{i=1}^{n_1} x_{i1}^2 + \sum_{i=1}^{n_2} x_{i2}^2 + \cdots + \sum_{i=1}^{n_k} x_{ik}^2 - \frac{T^2}{n}$$
$$= \sum_{i,j} x_{ij}^2 - \frac{T^2}{n}$$

+) Tổng bình phương do nhân tố ký hiệu là SSF (viết tắt của chữ Sumof Squares for Factor) được tính theo công thức sau:

$$SSF = \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$$
$$= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \cdots + \frac{T_k^2}{n_k} - \frac{T_2}{n}$$

+) Tổng bình phương do sai số ký hiệu là SSE (viết tắt của chữ Sumof Squares for the Error) được tính theo công thức:

$$SSE = \sum_{i=1}^{n_1} \left(x_{i1} - \bar{x}\right)^2 + \sum_{i=1}^{n_2} \left(x_{i2} - \bar{x}_2\right)^2 + \cdots + \sum_{i=1}^{n_k} \left(x_{ik} - \bar{x}_k\right)^2$$
$$= \sum_{i=1}^{n_1} x_{i1}^2 - \frac{T_1^2}{n_1} + \sum_{i=1}^{n_2} x_{i2}^2 - \frac{T_2^2}{n_2} + \cdots + \sum_{i=1}^{n_k} x_{ik}^2 - \frac{T_k^2}{n_k}$$
$$= \sum\sum x_{ij}^2 - \left(\frac{T_1^2}{n_1} + \cdots + \frac{T_k^2}{n_k}\right)$$

Từ công thức trên ta thấy

SST = SSF + SSE

+ Trung bình bình phương của nhân tố, ký hiệu là MSF (viết tắt của chữ Mean Square for Factor) được tính bởi công thức:

$$MSF = \frac{SSF}{k-1}$$

+ k – 1 được gọi là bậc tự do của nhân tố.

Trung bình bình phương của sai số, ký hiệu là MSS (viết tắt của chữ Mean Square for Error) được tính bởi công thức:

$$MSE = \frac{SSE}{n-k}$$

n – k được gọi là bậc tự do của sai số.

+ Tỷ số F được tính bởi công thức

$$F = \frac{MSF}{MSE}$$

Các kết quả nói trên được trình bày trong bảng sau đây gọi là

ANOVA (viết tắt của chữ Analysis of Variance: phân tích phương sai)

Bảng ANOVA

Nguồn	Tổng bình phương	Bậc tự do	Trung bình bình phương	Tỷ số F
Nhân tố	SSF	k – 1	MSF	MSF/MSE
Sai số	SSE	n – k	MSE	
Tổng số	SST	n – 1		

Người ta chứng minh được rằng nếu giả thiết H_o đúng thì tỷ số F

$$F = \frac{MSF}{MSE}$$

sẽ có phân bố Fisher với bậc tự do là (k – 1, n – k)

Thành thử giả thiết H_o sẽ bị bác bỏ ở mức ý nghĩa α của phân bố Fisher với bậc tự do là (k – 1, n – k). Trong bảng IV, k – 1 được gọi là bậc tự do ở mẫu số.

Phương pháp kiểm định nói trên được gọi là phân tích phương sai một nhân tố.

Cảm tưởng ban đầu của ta là ANOVA là một quá trình rất phức tạp. Nhưng thực ra nó khá đơn giản ngay cả khi ta chỉ có máy tính bỏ túi. Các bước trong ANOVA được tiến hành theo trình tự sau đây:

Bước 1: Tính SSF

Bước 2: Tính SST

Bước 3: Tính SSE = SST – SSF

Bước 4: Tính $MSF = \frac{SSF}{k-1}$

Bước 5: Tính $MSE = \frac{SSE}{n-1}$

Bước 6: Tính $F = \frac{MSF}{MSE}$

Bước 7: Tra bảng phân bố F để tìm c rồi so sánh với F và rút ra kết luận.

Ví dụ 5. thực hiện phân tích phương sai cho bảng số liệu sau đây.

	Các mức nhân tố				Tổng số
	1	2	3	4	
	12 10 7 8 9 14	12 16 15 9	9 7 16 11 7	12 8 8 10	
n_i	6	4	5	4	n = 19
T_i	60	52	40	38	T = 190

Bước 1.

$$\begin{aligned} SSF &= \frac{60^2}{6} + \frac{52^2}{4} + \frac{40^2}{5} + \frac{38^2}{4} - \frac{190^2}{19} \\ &= 1957 - 1900 = 57 \end{aligned}$$

Bước 2.

$$\begin{aligned} SST &= 12^2 + 10^2 + 7^2 + \cdots + 12^2 + 8^2 + 8^2 + 10^2 - \frac{190^2}{19} \\ &= 148 - 57 = 91 \end{aligned}$$

Bước 4.

$$MSF = \frac{SSF}{k-1} = \frac{57}{3} = 19$$

Bước 5.

$$MSE = \frac{SSE}{n-k} = \frac{148}{19-4} = \frac{148}{15} = 6,04$$

Bước 6.

$$F = \frac{MSF}{MSE} = \frac{19}{6,07} = 3,13$$

Ta trình bày các kết quả tính toán trên trong bảng ANOVA.

Nguồn	Tổng bình phương	Bậc tự do	Trung bình bình phương	Tỷ số F
Nhân tố	57	3	19	F = 3,13

Sai số	91	15	6,04
Tổng số	148	18	

Với mức ý nghĩa 5%, tra bảng phân bố Fisher với bậc tự do (3,15) ta được: c = 3,29.

Ta có F < c do đó ta chấp nhận H_o. ■

Ví dụ 6. Điểm thi của 12 sinh viên học các giáo sư A, B, C được cho trong bảng sau (thang điểm 100):

Giáo sư A	Giáo sư B	Giáo sư C
79	71	82
86	77	68
94	81	70
89	83	76

Với mức ý nghĩa 5%, kiểm định xem liệu điểm thi trung bình của các sinh viên theo học các giáo sư A, B, C có giống nhau hay không.

Giải. Kết quả tính toán cho ta bảng ANOVA như sau:

Nguồn	**Tổng bình phương**	**Bậc tự do**	**Trung bình bình phương**	**Tỷ số F**
Nhân tố	354,67	2	177,34	4,96
Sai số	322	9	35,78	
Tổng số	676,67	11		

Với mức ý nghĩa $\alpha = 5\%$, tra bảng phân bố Fisher với bậc tự do (2,9), ta tìm được c = 4,26.

Vì F > c nên ta bác bỏ H_o, nghĩa là điểm thi trung bình của các sinh viên theo học các giáo sư A, B, C là khác nhau ở mức ý nghĩa 5%.

Chú ý về sử dụng Minitab. Để tiến hành phân tích phương sai trên máy vi tính với phần mềm Minitab, đầu tiên ta nhập các số liệu vào dưới dạng các cột chẳng hạn các coat C_1, C_2, C_3, C_4.

Sau đó chỉ cần gõ lệnh

AOVONEWAY C1 – C4

là Minitab sẽ cho hiện lên màn hình bảng ANOVA tính trên dữ liệu đã đưa vào.

Ví dụ 7. Tiến hành phân tích phương sai bằng máy tính (sử dụng Minitab) bảng số liệu sau:

Điểm của các giáo sư			
An	**Vân**	**Ba**	**Bình**
56	61	58	68
64	66	60	74
67	52	65	59
61	48	49	54
70	47	75	66
	56		64

Giải

MTB > Mame C1 “An”

MTB > Mame C2 “Van”

MTB > Mame C3 “Ba”

MTB > Mame C4 “Binh”

MTB > Set C1

DATA > 56, 64, 67, 61, 70

DATA > End

MTB > Set C2

DATA > 61, 66, 52, 48, 47, 56

DATA > End

MTB > Set C3

DATA > 58, 60, 65, 79, 75

DATA > End

MTB > Set C4

DATA > 68, 74, 59, 54, 66, 64

DATA > End

MTB > AOVONEWAY C1 – C4

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
FACTOR	3	310,6	103,5	1,85	0,174
ERROR	18	1007,2	56,0		
TOTAL	21	1317,8			

Công việc còn lại là tra bảng phân bố Fisher với bậc tự do (3,18), mức α = 5% để tìm được c = 3, 16 số này nhỏ hơn F = 1,85. vậy ta chấp nhận H_o.

Giả sử việc phân tích phương sai dẫn tới bác bỏ H_o, nghĩa là có sự khác nhau giữa các trung bình. Như vậy tồn tại ít nhất một cặp μ_i, μ_j sao cho $\mu_i \neq \mu_j$. Đôi khi ta cần biết cụ thể cặp $\mu_i \neq \mu_j$ đó là cặp nào. Các nhà thống kê đã xây dựng được một số phương pháp để so sánh từng cặp giá trị trung bình hay so sánh những tổ hợp phức tạp hơn của các trung bình như phương pháp Dumcan, phương pháp Tukey, phương pháp Scheffe... Tuy nhiên trong giáo trình này ta không có điều kiện trình bày những phương pháp đó.

§ 4. PHÂN TÍCH PHƯƠNG SAI HAI NHÂN TỐ

Trên thực một biến lượng chịu tác động không chỉ một nhân tố mà có thể hai (hay nhiều nhân tố). Chẳng hạn năng suất cây trồng chịu ảnh hưởng của nhân tố giống và của nhân tố đất. Kết quả học tập của một sinh viên chịu ảnh hưởng không những bởi nhân tố giảng viên mà còn bởi nhân tố sĩ số của lớp học...

Trong mục này ta sẽ trình bày một cách vắn tắt kỹ thuật phân tích phương sai hai nhân tố nhằm phát hiện ảnh hưởng của mỗi nhân tố cũng như tác động qua lại của hai nhân tố đó đến biến lượng đang xét.

Giả sử chúng ta quan tâm tới nhân tố A và B. Nhân tố A được xem xét ở các mức A_1, A_2, ...A_r, và nhân tố B được xem xét ở các nước B_1, B_2,...B_c.

Gọi X_{jk} là ĐLNN đo lường hiệu quả việc tác động của mức A_j và B_k lên cá thể.

Giả sử $x_{1jk}, x_{2jk}, ..., x_{njk}$

là mẫu kích thước n_{jk} rút ra từ tập hợp chính các giá trị của X_{jk}. Ta gọi đó là mẫu (j, k). Ta đưa ra một số ký hiệu sau:

$\bar{x}_{jk}$: trung bình của mẫu (j, k)

$$n_{jo} = \sum_{k=1}^{c} n_{jk}$$

$$n_{ok} = \sum_{j=1}^{r} n_{jk}$$

$$n = \sum_{j} n_{jo} = \sum_{k} n_{ok}$$

$$\bar{x}_{jo} = \frac{\sum_{k} n_{jk}\bar{x}_{jk}}{n_{jo}} = \frac{\sum_{i}\sum_{k} x_{ijk}}{n_{jo}} = \text{ trung bình của mức } A_j$$

$$\bar{x}_{ok} = \frac{\sum_{j} n_{jk}\bar{x}_{jk}}{n_{ok}} = \frac{\sum_{i}\sum_{j} x_{ijk}}{n_{ok}} = \text{ trung bình của mức } B_k$$

$$\bar{x} = \text{trung bình chung} = \frac{\sum\sum\sum x_{jk}}{n} \bar{x}_{ok}$$

Ta có bảng sau đây ghi các kết quả tính toán trên:

A / B	B_1	B_2	...	B_k	...	B_c	Trung bình dòng A_j
A_1	$\bar{x}_{11}$	$\bar{x}_{12}$	...	$\bar{x}_{1k}$	...	$\bar{x}_{1c}$	$\bar{x}_{10}$
A_2	$\bar{x}_{21}$	$\bar{x}_{22}$	...	$\bar{x}_{2k}$	...	$\bar{x}_{2c}$	$\bar{x}_{20}$
...	...	...	...	...	...	...	...
A_j	$\bar{x}_{j1}$	$\bar{x}_{j2}$	...	$\bar{x}_{jk}$	...	$\bar{x}_{jc}$	$\bar{x}_{j0}$
...	...	...	...	...	...	...	...
A_r	$\bar{x}_{r1}$	$\bar{x}_{r2}$	...	$\bar{x}_{rk}$	...	$\bar{x}_{rc}$	$\bar{x}_{ro}$
Trung bình cột B_k	$\bar{x}_{o1}$	$\bar{x}_{o2}$	...		...	$\bar{x}_{oc}$	$\bar{x}$

+ Tổng bình phương chung, ký hiệu là SST, được tính theo công thức sau:

$$SST = \sum_{k=1}^{c}\sum_{j=1}^{r}\sum_{i=1}^{n_{jk}}\left(x_{ijk} - \bar{x}\right)^2$$

+ Tổng bình phương cho nhân tố A, ký hiệu là SSF_A được tính theo công thức sau:

$$SSF_B = \sum_{k=1}^{c} n_{ok} \left(\bar{x}_{ok} - \bar{x} \right)^2$$

+ Tổng bình phương do sai số, ký hiệu là SSE, được tính theo công thức

$$SSF = \sum_{k=1}^{c} \sum_{j=1}^{r} \sum_{i=1}^{n_{jk}} \left(x_{ijk} - \bar{x}_{jk}^{-2} \right)$$

+ Tổng bình phương do tương tác (Sum of Squares for Interaction) ký hiệu là SSI, được tính theo công thức.

$$SSI = \sum_{k=1}^{C} \sum_{j=1}^{r} \left(\bar{x}_{jk} - \bar{x}_{jo} - \bar{x}_{ko} + \bar{x} \right)^2$$

+ Trung bình bình phương của nhân tố A, ký hiệu là MSF_A, được tính bởi công thức:

$$MSF_A = \frac{SSF_A}{r-1}$$

r – 1 gọi là bậc tự do của A bằng số mức của A trừ 1.

+ Trung bình bình phương của nhân tố B, ký hiệu là MSF_B, được tính bởi công thức.

$$MSF_B = \frac{SSF_B}{c-1}$$

c – 1 gọi là bậc tự do của B bằng số mức của B trừ 1.

+ Trung bình bình phương của sai số, ký hiệu là MSE, được tính bởi

$$MSE = \frac{SSE}{n-cr}$$

n – cr gọi là bậc tự do của sai số.

+ Trung bình bình phương của tương tác, ký hiệu là MSI, được tính bởi

$$MSI = \frac{SSI}{(c-1)(r-1)}$$

(c – 1) (r – 1) gọi là bậc tự do của tương tác.

Chú ý rằng:

(r – 1) + (c – 1) + (c – 1) (r – 1) + n – rc = n – 1 = bậc tự do tổng cộng.

+ Tỷ số F cho nhân tố A, ký hiệu bởi F_A được tính như sau.

$$F_A = \frac{MSF_A}{MSE}$$

Tương tự tỷ số F cho nhân tố B, F_B được tính bởi

$$F_B = \frac{MSF_B}{MSE}$$

và tỷ số F cho tương tác giữa A và B, ký hiệu là F_{AB} được tính bởi:

$$F_{AB} = \frac{MSI}{MSE}$$

Với mức ý nghĩa α đã cho ta ký hiệu f (u, v) là phân vị mức α của phân bố Fisher với bậc tự do (u, v).

Ta có quy tắc quyết định như sau:

+ Nếu $F_A > f(r-1, n-cr)$ thì ta bác bỏ giả thiết.

H_o^A : "Các mức A_1,... Ar có hiệu quả trung bình như nhau"

+ Nếu $F_B > f(c-1, n-cr)$ thì ta bác bỏ giả thiết:

H_o^B : "Các mức B_1, B_2, ... B_c có hiệu quả trung bình như nhau"

Nếu $F_{AB} > f((r-1)(c-1), n-rc)$

Ta bác bỏ giả thiết:

H_o^{AB} : "Có sự tương tác giữa A và B".

Trên thực hành tính toán chúng ta thực hiện như sau:

Giả sử T_{jk} là tổng các giá trị trong mẫu (j, k). Ký hiệu

$$\begin{cases} T_{jo} = \sum_{k=1}^{c} T_{jk}, \quad T_{ok} = \sum_{j=1}^{r} T_{jk} \\ n_{jo} = \sum_{k=1}^{c} n_{jk}, \quad n_{ok} = \sum_{j=1}^{r} n_{jk} \end{cases}$$

$$\begin{cases} T = \sum T_{jo} = \sum T_{ok} = \sum\sum\sum x_{ijk} \\ n = \sum n_{jo} = \sum n_{ok} \end{cases}$$

$$A = \sum\sum\sum x_{ijk}^2 \qquad (3)$$

Ta có các đẳng thức sau:

$$SST = A - \frac{T^2}{n} \tag{4}$$

$$SSF_A = \sum_{j=1}^{r} \frac{T_{jo}^2}{n_{jo}} - \frac{T^2}{n} \tag{5}$$

$$SSF_B = \sum_{k=1}^{c} \frac{T_{ok}^2}{n_{ok}} - \frac{T^2}{n} \tag{6}$$

$$SSE = A - \sum_{k=1}^{c} \sum_{j=1}^{r} \frac{T_{jk}^2}{n_{jk}} \tag{7}$$

$$SSI = SST - SSF_A - SSF_B - SSE \tag{8}$$

Đặc biệt nếu tất cả các mẫu bằng nhau $n_{jk} = m$ với mọi j, k thì:

$$n_{jo} = cm,\ n_{ok} = rm$$

do đó

$$SSF_A = \frac{\sum_{j=1}^{r} T_{jo}^2}{cm} - \frac{T^2}{n} \tag{5'}$$

$$SSF_B = \frac{\sum_{k=1}^{r} T_{ok}^2}{rm} - \frac{T^2}{n} \tag{6'}$$

$$SSE = A - \frac{\sum_{k} \sum_{j} T_{jk}^2}{m} \tag{7'}$$

Trước hết ta cần tính các đại lượng T_{jk}. Tiếp theo tính các giá trị T_{jo}, n_{jo}, n_{ok}, T_{ok}, n, T và A theo các công thức (1), (2), (3).

Từ đó tính SST, SSF_A, SSF_B, SSE và SSI theo các công thức (4), (5), (6), (7) (hoặc (5'), (6'), (7') nếu $n_{jk} = m$).

PHÂN TÍCH TƯƠNG QUAN VÀ HỒI QUY

§ 1 PHÂN TÍCH TƯƠNG QUAN TUYẾN TÍNH

Giả sử X và Y là hai biến lượng (hay còn gọi là hai ĐLNN). Chúng ta đã biết rằng X và Y được gọi là độc lập nếu việc ĐLNN này nhận một giá trị nào đó (bất kỳ) cũng không ảnh hưởng gì đến phân bố xác suất của ĐLNN kia. Tuy nhiên trong nhiều tình huống thực tế, X và Y không độc lập với nhau. Điều này thường gặp, chẳng hạn khi X và Y là hai ghép đo nào đó tiến hành trên cùng một cá thể. Ví dụ X là chiều dài cánh tay Y là chiều cao của một người; hoặc X là điểm thi tốt nghiệp tú tài và Y là điểm thi vào đại học của cùng một học sinh.

Để đo mức độ phụ thuộc tuyến tính giữa hai ĐLNN X và Y, người ta đưa ra khái niệm hệ số tương quan. Hệ số tương quan lý thuyết của X và Y, ký hiệu là ρ, được định nghĩa bởi công thức

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

ở đó μ_X, σ_X là giá trị trung bình và độ lêchhj tiêu chuẩn của X, và μ_Y, σ_Y là giá trị trung bình và độ lệch tiêu chuẩn của Y.

Người ta đã chứng minh được ρ là một số nằm trong giai đoạn $[-1,1]$. Khi $\rho = 0$ thì không có tương quan tuyến tính giữa X và Y. Đặc biệt nếu (X,Y) có phân bố chuẩn thì $\rho = 0$ khi và chỉ khi X, Y độc lập. Khi $|\rho|$ càng gần 1 thì sự phụ thuộc tuyến tính giữa X và Y càng mạnh. Nếu $|\rho| = 1$ thì Y thì một hàm tuyến tính cảu X.

Muốn biết được ρ chúng ta cần biết phân bố của tập hợp chính bao gồm tất cả các giá trị của cặp (X, Y). Tuy nhiên thông tin này thường là khó nắm bắt.

Vì vậy, tương tự như vấn đề ước lượng và kiểm định giá trị trung bình hay phương sai đã xét ở các chương trước, chúng ta có bài toán ước lượng và kiểm định hệ số tương quan ρ căn cứ trên một mẫu quan sát (x_1, y_1) (x_1, y_2),..., (x_n, y_n) các giá trị của (X, Y).

Đại lượng sau đây được sử dụng như một ước lượng cho ρ:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

r được gọi là hệ số tương quan.

Để tính toán cho thuận lợi, r có thể viết dưới dạng sau:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Nên nhớ rằng r cũng nằm trong đoạn [–1,1]. Vì vậy nếu thu được giá trị r nằm ngoài đoạn [–1,1] có nghĩa là ta đã tính toán sai.

Ví dụ 1. Tính hệ số tương quan r dựa trên mẫu gồm 10 quan sát sau đây:

(80; 2,4) ; (85 ; 2,8) ; (88 ; 3,3) ; (90 ; 3,1) ; (95 ; 3,7) ; (92 ; 3); (82 ; 2,5) ; (75 ; 2,3) ; (78 ; 2,8) ; (85 ; 3,1).

Giải. Đầu tiên ta hãy tính các tổng $\sum x$, $\sum y$, $\sum xy$, $\sum x^2$, $\sum y^2$. Điều này có thể thực hiện dễ dàng bằng máy tính bỏ túi.

Ta có $\sum xy = 2486,3$; $\sum x = 850$;

$\sum x^2 = 72617$; $\sum y = 29$;

$\sum y^2 = 85,78$.

Vậy $n\sum xy - (\sum x).(\sum y) = 10(2486,3) - (850)(29)$

$= 24863 - 24650 = 213$

$n(\sum x^2) - (\sum x)^2 = 10(72617) - (850)^2$

$= 726170 - 722500 = 3670$

và $n(\sum y^2) - (\sum y)^2 = 10(85,78) - 29^2$

$= 857,8 - 841 = 18,8.$

Vậy hệ số tương quan r là

$$r = \frac{n\Sigma xy - (\Sigma x).(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{213}{\sqrt{3670}.\sqrt{16,8}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}.$$

■

Nếu có phần mềm Minitab ta sẽ tính hệ số tương quan chỉ bằng một lệnh đơn giản

CORRELATION C2 C1

Trong đó có hai dãy số liệu (x_1) (y_1) được nhập tương ứng vào các cột C1 và C2.

Ví dụ 2. Một nhà nghiên cứu quan tâm tới mối liên hệ giữa tuổi và mạch đập của phụ nữ. Trong một mẫu quan sát gồm 5 phụ nữ chọn được ngẫu nhiên có số liệu sau, ở đó X là tuổi, Y là nhịp mạch đập.

	X	**Y**	**XY**	$\mathbf{X^2}$	$\mathbf{Y^2}$
	23	210	4830	529	44100
	39	185	7215	1521	34255
	19	220	4180	361	48400
	44	164	7216	1936	26896
	51	123	6273	2601	15129
Tổng	176	902	29714	6948	168750

Nếu tính bằng ta thì

$$n\Sigma xy - (\Sigma x).(\Sigma y) = 5(29174) - (176).(902)$$

$$= 148570 - 158752 = -10182$$

$$n\Sigma x^2 - (\Sigma x)^2 = 34740 - 30976 = 3764$$

$$n\Sigma y^2 - (\Sigma y)^2 = 843750 - 813604 = 30146$$

$$r = \frac{10182}{\sqrt{3764}\,\sqrt{30146}} = \frac{10182}{(61,35).(173,62)} = -0,956$$

Nếu sử dụng Minitab ta sẽ gõ các lệnh sau

```
MTB > SET C1
DATA > 23  39  19  44  51
DATA > END
MTB > SET  C2
DATA > 210  185   220   164   123
DATA > END
MTB > CORRELATION  C1  C2
```

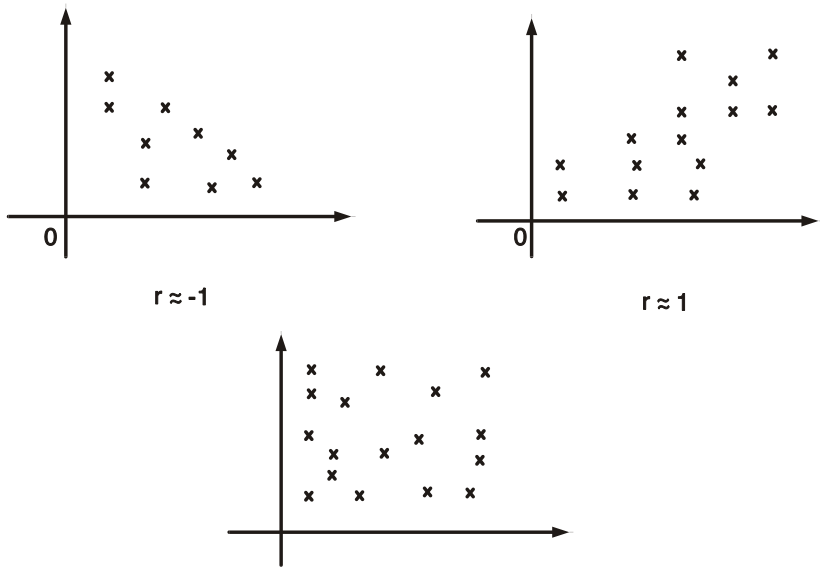
Sau đó màn hình sẽ hiện ra

Correlation of C1 and C2 = –0,956.

Để có một khái niệm sơ bộ về mối quan hệ giữa các ĐLNN X và Y trước tính hệ số tương quan người ta thường biểu diễn mỗi quan sát (x_i,y_1) bởi một điểm trên mặt phẳng với các tọa độ là (x_i,y_i). Giả sử ta có n quan sát (x_1,y_1), (x_2,y_2),..., (x_n,y_n).

Chúng được biểu diễn thành một tập hợp điểm trên mặt phẳng gọi là đám mây điểm. Nếu các điểm này có xu hướng tụ tập xung quanh một đường thẳng nào đó thì hệ số tương quan r có trị tuyệt đối khá gần 1, còn nếu nó nằm rải rác thành một hình tròn (đám mây điểm tròn hoặc vuông) thì |r| rất gần 0.

Các hình vẽ dưới đây minh họa các trường hợp $r \approx -1$



Khi sử dụng Minitab ta cần đánh lệnh

PLOT C2 C1

trong đó ta nhập các dữ liệu x_1 vào cột C1 còn các dữ liệu y_i vào cột C2. Màn hình sẽ cung cấp ngay cho ta một đám mây điểm.

Tiếp theo chúng ta đề cập vấn đề kiểm định giả thiết về hệ số tương quan lý thuyết ρ của tập hợp chính (bao gồm toàn bộ các quan sát có thể của (X, Y)). Kiểm định đầu tiên và quan trọng nhất là kiểm định xem X và Y có tương quan với nhau không. Chúng ta có bài toán kiểm định.

H_o: $\rho = 0$ (X, Y không tương quan)

Với đối thiết $H_1: \rho \neq 0$

Việc xây dựng quy tắc kiểm định bài toán trên dựa vào định lý sau.

Định lý. Nếu (X, Y) có phân bố chuẩn hai chiều thì dưới giả thiết H_o, ĐLNN

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

có phân bố Student với n –2 bậc tự do.

Thành thử test thống kê thích hợp cho bài toán kiểm định này là

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Ta sẽ bác bỏ H_o nếu $|T| > c$, ở đó c là phân vị mức $\frac{\alpha}{2}$ của phân bố Student với n – 2 bậc tự do.

Ví dụ 3. Trong một mẫu gồm 42 quan sát (x_i, y_i) rút ra từ tập hợp chính các giá trị của (X, Y), chúng ta tính được hệ số tương quan mẫu là r = 0,22. Với mức ý nghĩa $\alpha = 5\%$, có thể kết luận rằng X và Y có tương quan hay không?

Giải. Ta có $T = \frac{0,22\sqrt{40}}{\sqrt{1-(0,22)^2}} = \frac{0,22}{0,154} = 1,43$

Với bậc tự do 40, $\alpha = 5\%$, ta tìm được hàng số c là 2,021.

Vậy ta chưa có cơ sở bác bỏ H_o, nghĩa là chưa kết luận được X và Y có tương quan.

Với bài toán kiểm định giả thiết

$H_o: \rho = \rho_o$

$H_1: \rho \neq \rho_o$

ở đó ρ_o là một giá trị khác không cho trước, ta sẽ xây dựng test thống kê

$$T = \frac{u-m}{\sigma}$$

ở đó $u = \frac{1}{2}\ln\frac{1+r}{1-r}$

$$m = \frac{1}{2}\ln\frac{1+\rho_o}{1-\rho_o}$$

$$\sigma = \frac{1}{\sqrt{n-3}}.$$

Người ta đã chứng minh được rằng nếu giả thiết H_o đúng thì T sẽ có phân bố xấp xỉ phân bố chuẩn tắc N(0,1). Thành thử H_o sẽ bị bác bỏ ở mức ý nghĩa α nếu $|T| > c$, trong đó c là phân vị mức $\frac{\alpha}{2}$ của phân bố chuẩn tắc.

Ví dụ 4. Từ một mẫu kích thước n = 35 rút ra từ tập hợp chính các giá trị của (X, Y), ta tính được hệ số tương quan là r = 0,8. Với mức ý nghĩa $\alpha = 5\%$, kiểm định giả thiết

$$H_o : \rho = 0,9$$

$$H_1 : \rho \neq 0,9$$

Giải. Ta có $$u = \frac{1}{2}\ln\frac{1+0,8}{1-0,8} = 1,009$$

$$m = \frac{1}{2}\ln\frac{1+0,9}{1-0,9} = 1,472$$

$$\sigma = \frac{1}{\sqrt{32}} = 0,177$$

Từ đó $T = \frac{1,099 - 1,472}{0,177} = -2,11$

Với $\alpha = 5\%$, ta tìm được c = 1,96

Vì $|T| = 2,11 > 1,96$, nên ta bác bỏ H_o, nghĩa là $\rho \neq 0,9$. ■

Test thống kê nói trên $T = \frac{u-m}{\sigma}$ cũng cho phép ta xác định được khoảng tin cậy cho hệ số tương quan lý thuyết ρ.

Ví dụ 5. Trong một mẫu kích thước n = 52 rút ra từ tập hợp chính các giá trị của (X,Y), ta tính được hệ số tương quan là r = 0,53. Căn cứ trên kết quả đó hãy cho một khoảng tin cậy 95% cho hệ số tương quan lý thuyết ρ_o giữa X và Y.

Giải. Ta có $$u = \frac{1}{2}\ln\frac{1+0,53}{1-0,53} = 0,59$$

$$\sigma = \frac{1}{49} = \frac{1}{7} = 0,43$$

Vì $T = \frac{u - m}{\sigma}$ có phân bố chuẩn tắc, do đó với c là phân vị mức $\frac{\alpha}{2}$ của phân bố chuẩn tắc N(0,1), ta có

$$P\{|T| < c\} = 1 - \alpha.$$

Với $1 - \alpha = 0,95$ suy ra $\alpha = 0,05$, ta có $c = 1,96$.

Vậy với xác suất 0,95 ta có

$$-c\sigma < u - m < c\sigma$$

$$\Leftrightarrow u - c\sigma < m < m + c\sigma$$

Thay giá trị của u, c, σ vào ta được $0,31 < m < 0,87$

hay
$$0,31 < \frac{1}{2}\ln\frac{1+\rho_o}{1-\rho_o} < 0,87$$

$$\Leftrightarrow \quad 0,62 < \ln\frac{1+\rho_o}{1-\rho_o} < 1,74$$

$$\Leftrightarrow \quad e^{0,62} < \frac{1+\rho_o}{1-\rho_o} < e^{1,74}$$

$$\Leftrightarrow 1,858 < \frac{1+\rho_o}{1-\rho_o} < 5,7$$

Từ bất đẳng thức trên dễ dàng tìm được

$$0,3 < \rho_o < 0,7.$$

Đó là khoảng tin cậy cho ρ_o. ■

Cuối cùng ta cần lưu ý một số điểm sau.

Chú thích.

1) Hệ số tương quan chỉ là một số đo mối quan hệ tuyến tính giữa X và Y.

2) Nếu X và Y độc lập thì hệ số tương quan giữa chúng bằng 0. Điều ngược lại chưa chắc đúng (trừ khi X và Y có phân bố chuẩn đồng thời).

Có thể xảy ra trường hợp X và Y không tương quan ($\rho = 0$) nhưng Y lại là một hàm của X (tức là giữa X và Y có sự phụ thuộc hàm).

3) Mối quan hệ tuyến tính được đo bởi hệ số tương quan hoàn toàn chỉ là một chỉ số toán học. Nó có thể không biểu thị một mối quản hệ nhân quả nào.

Hệ số tương quan của X và Y có thể rất cao chỉ vì chúng đều liên quan tới một biến thứ ba.

Ví dụ. Tính toán trên các số liệu thống kê từ năm 1961 đến năm 1977 ở Mỹ cho thấy hệ số tương quan giữa lương của giáo viên và giá bán của rượu là rất cao. Rõ ràng chúng ta không thể cho rằng tăng giá rượu (hay giảm) sẽ làm tăng (hay giảm) lương giáo viên, hay tăng lương (hay giảm lượng) giáo viên sẽ kéo theo tăng hay giảm giá rượu.

Để giải thích hiện tượng này ta cần tìm một nhân tố thứ ba, nhân tố này sẽ là nguyên nhân của việc tăng lượng và tăng giá rượu.

Nhân tố đó chính là sự lạm phát. Lạm phát đã dẫn đến việc phải tăng lương cho giáo viên và tăng giá rượu. Như vậy sự tương quan cao giữa tiền lương giáo viên và giá rượu chỉ đơn thuần phản ánh một hiệu ứng chung của việc gia tăng theo gần như cùng một nhịp của hai biến đó.

Ví dụ. Các số liệu thống kê vào cuối những năm 1800 cho thấy có một sự tương quan cao giữa số con cò và số trẻ mới sinh trong các thành phố của châu Âu. Thật là ngớ ngẩn nếu cho rằng số cò và số trẻ sơ sinh có mối quan hệ nhân quả. Cách giải thích đúng đắn hiện tượng này là trong thời gian đó, thành phố được phát triển bởi nhiều nhà có mái tranh. Mai tranh lại là nơi trú ngụ lý tưởng cho các con cò.

Thành thử có nhiều nhà có mái tránh sẽ thu hút nhiều cò và mặt khác nhiều nhà tức là nhiều gia đình, dĩ nhiêu sẽ sinh ra nhiều đứa trẻ.

Tóm lại sự giả thích đúng đắn lý do của sự tương quan giữa hai biến X và Y đòi hỏi một kiến thức tổng hợp đôi khi nằm ngoài Toán học và Thống kê.

§ 2. KIỂM TRA TÍNH ĐỘC LẬP

Giả sử ta quan tâm tới một dấu hiệu nào đó của các cá thể trong một tập hợp chính C. Dấu hiệu này nói chung thay đổi từ cá thể này sang cá thể khác. Nếu dấu hiệu này biểu thị được bởi một con số, hay nói cách khác có thể gán số đo cho dấu hiệu này lên các cá thể, thì ta nói dấu hiệu này là một biến lượng hay là một dấu hiệu định lượng. Chẳng hạn nếu cad thể là người thì biến lượng có thể là chiều cao, trong lượng, tuổi... tuy nhiên trong thực tế có những dấu hiệu không thể đo đạc để biểu diễn bằng con số được. Chẳng hạn màu tóc, màu mắt của một người, cảm giác hạnh phúc, sự yêu thích một cuốn phim nào đó...

Đó đều là những dấu hiệu không đo đạc được. Ta gọi đó là những dấu hiệu định tính.

Trong mục này ta sẽ xét bài toán kiểm tra tính độc lập của hai dấu hiệu. Trước hết, chúng ta xét bài toán kiểm định tính độc lập của dấu hiệu định tính A và B.

Ta chia dấu hiệu A ra làm r mức độ A_1, A_2,..., A_r, và chia đặc tính B làm k mức độ B_1, B_2,..., B_k. Xét một mẫu ngẫu nhiên gồm n cá thể. Mỗi cá thể sẽ mang dấu hiệu A ở mức A_i nào đó và mang dấu hiệu B ở mức B_j nào đó. Giả sử n_{ij} là số cá thể có các dấu hiệu A_i và B_j. Các số liệu n_{ij} được ghi trong bảng sau đây gọi là bảng liên hợp các dấu hiệu (Contingency Table).

A / B	B_1	B_2	...	B_k	Tổng
A_1	n_{11}	n_{12}	...	n_{1k}	n_{10}
A_2	n_{21}	n_{22}	...	n_{2k}	n_{20}
...	...	...	...	...	...
A_r	n_{r1}	n_{r2}	...	n_{rk}	n_{r0}
Tổng	n_{01}	n_{02}	...	n_{0k}	n

Trong đó ký hiệu p_{ij} là xác suất để một cá thể chọn ngẫu nhiên mang dấu hiệu A_i và B_j ; p_{jo} và p_{oj} tương ứng là xác suất để cá thể mang dấu hiệu A_i và B_j.

Nếu giả thiết H_o "Hai dấu hiệu A và B độc lập" chúng ta có hệ thức sau:

$$p_{ij} = p_{io}.p_{oj}$$

Các xác suất p_{io} và p_{oj} được ước lượng bởi

$$p_{io} \approx \frac{n_{io}}{n},$$

$$p_{oj} \approx \frac{n_{oj}}{n}$$

Do đó H_o đúng thì

$$p_{ij} \approx p_{io}.p_{oj} = \frac{n_{io}.n_{oj}}{n^2},$$

và số cá thể có đồng thời dấu hiệu A_i và B_j sẽ xấp xỉ bằng

$$n_{ij} = np_{ij} = \frac{n_{io}n_{oj}}{n}$$

Các số n_{ij} được gọi là các tần số lý thuyết (TSLT), còn các số n_{ij} được gọi là các tần số quan sát (TSQS). Khoảng cách giữa các TSLT và TSQS được đo bằng đại lượng sau:

$$T = \sum_{j=1}^{k}\sum_{i=1}^{r}\frac{(n_{ij} - n_{ij})^2}{n_{ij}}$$

Người ta đã chứng minh được rằng nếu n lớn và các TSLT không nhỏ hơn 5 thì T sẽ có phân bố xấp xỉ phân bố χ^2 với bậc tự do là (k–1).(r–1). Thành thử H_o sẽ bị bác bỏ ở mức ý nghĩa α nếu T > c, trong đó c là phân vị mức α của phân bố χ^2 với (k–1).(r–1) bậc tự do.

Chú ý. Ta có các thức sau đây khá thuận lợi trong tính toán thực hành:

$$T = n\left\{\sum\frac{n_{ij}^2}{n_{io}n_{oj}} - 1\right\}$$

Trong trường hợp k = r = 2 (bảng liên hợp có hai dòng, hai cột) thì

$$T = \frac{n\begin{vmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{vmatrix}}{n_{01}n_{02}n_{10}n_{20}}$$

trong đó $\begin{vmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{vmatrix} = n_{11}n_{22} - n_{21}n_{22}$

là định thức của ma trận $\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$.

Ví dụ 6. Ở các cây ngọc trâm lá hai dạng "lá phẳng" hoặc "lá nhãn", hoa có hai dạng "hoa bình thường "hoặc" hoa hoàng hậu". Quan sát một mẫu gồm 560 cây ngọc trâm ta thu được kết quả sau:

Hoa / Lá	Bình thường	Hoàng hậu	Tổng số
Phẳng	328	122	450
Nhãn	77	33	110

Tổng số	405	155	560

Có thể chấp nhận hai đặc tính về hoa và lá nói trên là độc lập hay không? Hay là giữa chúng có sự liên kết?

Giải. Ta có

$$T = \frac{560\begin{vmatrix} 328 & 122 \\ 77 & 33 \end{vmatrix}}{(450).(110).(405).(155)} = 0,368$$

Với mức ý nghĩa 5%, tra bảng phân bố χ^2 với bậc tự do ta tìm được c = $\chi^2_{0,05}$ = 3,841.

T nhỏ hơi c, vậy ta chấp nhận giả thiết: Hai đặc tính về hoa và là nói trên độc lập. ■

Tiêu chuẩn χ^2 nói trên còn có thể áp dụng để kiểm định tính độc lập của một dấu hiệu định tính A và một dấu hiệu định lượng (biến lượng) X. Khi đó ta cần chia miền giá trị của X thành k khoảng B_1, B_2,..., B_k và nếu cá thể có số đo x_i rơi vào khoảng B_j thì ta xem như cá thể đó có dấu hiệu B_j.

Tương tự như vậy ta có thể dùng tiêu chuẩn χ^2 nói trên để kiểm tra tính độc lập của hai ĐLNN X và Y (Lưu ý rằng nếu X và Y không tương quan thì chưa chắc X và Y đã độc lập). Muốn vậy ta cần chia miền giá trị của X thành k khoảng B_1, B_2,..., B_k còn miền giá trị của Y thành r khoảng A_1,..., A_r. Nếu cá thể có số đo (x,y) trong đó $x \in B_i$, $y \in A_j$, thì ta coi như cá thể đó có các dấu hiệu B_i và A_j.

Ví dụ 7. Một con ốc sên rừng có thể có màu vỏ là vàng hoặc hồng. Số vạch trên vỏ của nó có thể là 0, 1, 2, 3, 4, 5.

Ở đây dấu hiệu A (màu đỏ) là dấu hiệu định tính với hai mức vàng, hồng còn số vạch trên vỏ X là một dấu hiệu định lượng (hay X là một ĐLNN rời rạc). Ta muốn kiểm định xem A và X có độc lập hay không.

Giải. Ta chia tập giá trị của X làm các mức

B_1 = {không có vạch}

B_2 = {1 hay 2 vạch}

B_3 = {3 hay 4 vạch}

B_4 = {5 vạch}

Xét một mẫu ngẫu nhiên gồm 169 con ốc sên ta, thu được số liệu sau đây.

Số vạch / Màu đỏ	B_1	B_2	B_3	B_4	Tổng số
Vàng	35	19	36	25	115
Hồng	14	14	16	10	54
Tổng số	49	33	52	35	169

Ta có $T = 169\left\{\frac{35^2}{(49).(115)} + \frac{19^2}{(33).(115)} + ... + \frac{10^2}{(35).(54)} - 1\right\} = 2,13$

Với mức ý nghĩa $\alpha = 5\%$ tra bảng phân bố χ^2 với bậc tự do là $(2-1).(4-1) = 3$, ta tìm được $c = \chi^2_{0,05} = 7,81$. Ta có $T < c$ vậy giả thiết H_o phù hợp với số liệu thực nghiệm. Ta chấp nhận rằng A và X độc lập.

Ví dụ 8. Giả sử X và Y tương ứng là số đo huyết áp và trọng lượng (tính bằng pound) (1 pound = 0,454 kg) của trẻ em 14 tuổi. Ta muốn khẳng định xem X và Y có độc lập không.

Giải. Chia X thành các mức

$B_1 = \{X \le 99\}$;

$B_2 = \{99 < X \le 110\}$;

$B_3 = \{110 < X \le 120\}$;

$B_4 = \{X > 120\}$.

Chia Y làm hai mức

$A_1 = \{Y \le 102\}$;

$A_2 = \{Y > 102\}$.

Một mẫu gồm 200 trẻ em được đo huyết áp và trọng lượng cho thấy số liệu sau:

Huyết áp / Trọng lượng	B_1	B_2	B_3	B_4	Tổng số
A_1	10	20	11	5	46
A_2	6	48	50	50	154

Tổng số	16	68	61	55	200

Ta có: $T = 200\left\{\frac{10^2}{(16)(46)} + \frac{20^2}{(68)(46)} + \cdots + \frac{50^2}{(55)(154)} - 1\right\} = 22,53$

Với mức ý nghĩa $\alpha = 1\%$, tra bảng phân bố χ^2 với bậc tự do là (2 – 1)(4 – 1) = 3, ta tìm được $c = \chi^2_{0,01} = 11,345$. Vì T > c nên ta bác bỏ H_o và kết luận:

Giữa huyết áp và trọng lượng trẻ 14 tuổi có sự phụ thuộc lẫn nhau.

* § 3. PHÂN TÍCH TƯƠNG QUAN PHI TUYẾN

Như đã nói trong §1, hệ số tương quan dùng để đo mức độ phụ thuộc tuyến tính giữa hai ĐLNN. Như thế chúng ta còn chưa có một chỉ tiêu để đo mức độ phụ thuộc nói chung. Cần nhớ rằng nếu hệ số tương quan giữa X và Y rất bé hay thậm chí bằng 0 thì giữa X và Y vẫn có thể có một mối liên hệ phi tuyến rất chặt chẽ.

Để đo mức độ phụ thuộc nói chung của ĐLNN Y vào ĐLNN X, người ta đưa ra khái niệm tỷ số tương quan. Tỷ số tương quan lý thuyết của Y theo X được ký hiệu bởi $\eta^2_{Y/X}$ là một số không âm xác định theo công thức sau đây.

$$\eta^2_{Y/X} = 1 - \frac{E(Y - E(Y/X))^2}{DY} = \frac{DY - E(Y - E(Y/X))^2}{DY}$$

trong đó E[Y/X] ký hiệu kỳ vọng của Y tính trong điều kiện X cố định một giá trị. E[Y/X] gọi là kỳ vọng của Y với điều kiện X.

Người ta đã chứng minh được rằng

$$0 \le \eta^2_{Y/X} \le 1 \text{ và } \rho^2 \le \eta^2_{Y/X}$$

Hiệu số $\eta^2_{Y/X} - \rho^2$ đo mức độ phụ thuộc phi tuyến giữa Y và X.

Nếu hiệu số $\eta^2_{Y/X} - \rho^2$ càng lớn thì có nghĩa là có sự tương quan phi tuyến càng mạnh.

Bay giờ ta xét vấn đề ước lượng và kiểm định giả thiết về tỷ số tương quan. Giả sử (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) là một mẫu gồm n quan sát độc lập rút ra từ tập hợp chính các giá trị của (X, Y). Chúng ta cần giả thiết rằng trong dãy các giá trị của X: x_i, x_2,...x_n, mỗi giá trị x_i đều được lặp lại ít nhất một lần. Giả sử $x_{(1)} < x_{(2)} ... < x_{(k)}$ là các giá trị khác nhau trong dãy (x_i). Ta sẽ trình bày dãy số liệu (x_i, y_i) thành bảng sau đây, được gọi là bảng tương quan.

X / Y	$x_{(1)}$	$x_{(2)}$	...	$x_{(k)}$	
	y_{11} y_{21} ... $y_{n_1 1}$	y_{12} y_{22} ... $y_{n_2 2}$	... ...	y_{1k} y_{2k} ... $y_{n_k k}$	
	n_1	n_2	...	n_k	$n = \sum n_i$
	T_1	T_2	...	T_k	$n = \sum T_i$

Bảng này rất giống với bảng số liệu khi tiến hành phân tích phương sai (xem chương VI, § 3).

Tiếp theo ta tiến hành phân tích phương sai.

Ký hiệu: $T_i = \sum_{i=1}^{n_i} y_{ji}$ (tổng các số liệu y_{ji} ở cột $x_{(i)}$)

$$T = \sum T_i$$

n_i là số các số liệu ở cột $x_{(i)}$ (cũng chính là số các giá trị x_j mà $x_j = x_{(i)}$)

Nhớ lại rằng (xem chương VI, § 3):

+ Tổng bình phương chung SST được tính bởi công thức:

$$STT = \sum\sum y_{ij}^2 - \frac{T^2}{n}$$

+ Tổng bình phương do nhân tố SSF được tính bởi công thức

$$SSF = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

Đại lượng sau đây được sử dụng như là một ước lượng cho tỷ số tương quan lý thuyết $\eta^2_{Y/X}$:

$$\eta^2_{Y/X} = \frac{SSF}{SST}$$

$\eta^2_{Y/X}$ được gọi là tỷ số tương quan của Y đối với X. Để cho gọn từ nay ta sẽ viết η^2 thay cho $\eta^2_{Y/X}$.

Người ta đã chứng minh được rằng

$$0 \le r^2 \le \eta^2$$

ở đó r là hệ số tương quan. Bình phương của hệ số tương quan r^2 được gọi là hệ số xác định.

Tỷ số tương quan η^2 được lý giải như là tỷ lệ biến động của Y do có sự phụ thuộc của Y vào X.

Hệ số xác định r^2 được lý giải như là tỷ lệ biến động của Y do có sự phụ thuộc tuyến tính của Y vào X.

Ví dụ 9. Cho mẫu quan sát sau đây của cặp ĐLNN (X, Y):

(8, 82); (8, 78); (12, 65); (12, 50); (20, 60); (20, 47); (24, 52); (24, 41); (8, 87); (8, 58); (8, 70); (12, 62); (12, 55); (12, 52); (20, 44); (20, 66); (20, 41); (24, 57); (24, 50); (24, 47); (8, 65); (12, 49); (20, 57); (24, 65).

Hãy tính hệ số tương quan hệ số xác định và tỷ số tương quan của Y đối với X.

Giải. Trước hết ta cần trình bày các số liệu trên dưới dạng bảng tương quan sau đây:

X / Y	8	12	20	24	
	82 78 87 58 70 65	65 50 62 55 52 49	60 47 44 66 41 57	52 41 57 50 63	
n_i	6	6	6	6	n = 24
T_i	440	333	315	310	T = 1398

+ Tính hệ số tương quan

Ta có

$$\sum x = 6(8) + 6(12) + 6(20) + 6(24) = 384;$$
$$\sum y = T = 1398;$$
$$\sum x^2 = 6(64) + 6(144) + 6.(400) + 6.(576) = 7104$$
$$\sum y^2 = 82^2 + 78^2 + \cdots + 63^2 = 84908;$$
$$\sum xy = 8(440) + 12(333) + 20(315) + 24(310) = 21256$$

Vậy $n\sum xy - \left(\sum x\right)\left(\sum y\right) = -26688;$

$$\sqrt{n\sum x^2 - \left(\sum x\right)^2} = \sqrt{24(7104) - 384^2} = 151,789;$$
$$\sqrt{n\sum y^2 - \left(\sum y\right)^2} = \sqrt{24(84909) - 1398^2} = 288,77.$$

Thành thử $r = \dfrac{-26688}{(151,789)(288,77)} = 0,6089$

Hệ số xác định là $r^2 = 0,6089^2 = 0,37$

+ Tính tỷ số tương quan

Ta có:

$$SST = \sum y^2 - \frac{T^2}{n} = 84908 - \frac{1398^2}{24} = 3474,5;$$
$$SSF = \sum \frac{T_i^2}{n_i} - \frac{T^2}{n} = \frac{440^2 + \cdots + 310^2}{0} - \frac{1398^2}{24} = 1868,83$$

Từ đó $\eta^2 = \dfrac{SSF}{SST} = 0,5378$ ■

Hiệu số $\eta^2 - \rho^2$ giữa tỷ số tương quan lý thuyết và hệ số xác định lý thuyết cho ta một hình ảnh về sự phụ thuộc phi tuyến của Y đối với X. Nếu hiệu số đó bằng 0 thì điều đó nghĩa là chỉ có tương quan tuyến tính giữa Y và X.

Để kiểm định giả thiết

H_o: $\eta^2 - \rho^2 \neq 0$ (không có tương quan phi tuyến), với đối thiết

H_1: $\eta^2 - \rho^2 > 0$ (có tương quan phi tuyến), ta dùng test thống kê sau:

$$F = \frac{\frac{\eta^2 - r^2}{k-2}}{\frac{1-\eta^2}{n-k}} = \frac{\left(\eta^2 - r^2\right)(n-k)}{\left(1-\eta^2\right)(k-2)}$$

Người ta đã chứng minh được rằng nếu H_o đúng thì F sẽ có phân Fisher với bậc tự do là (k – 2, n – k). Thành thử giả thiết H_o: "Không có tương quan phi tuyến" sẽ bị bác bỏ ở mức α nếu F lớn hơn hằng số c là phân vị mức α của phân bố Fisher với bậc tự do là (k – 2, n – k).

Ví dụ 10. Trở lại ví dụ trên ta muốn kiểm tra xem liệu có tương quan phi tuyến của Y đối với X hay không.

Ta có $F = \frac{(0,5378 - 0,37)}{(1-0,5378)} \frac{(24-4)}{(4-2)} = \frac{(0,1678)(20)}{(0,4622).2} = 3,63$

Tra bảng phân bố Fisher với bậc tự do (2, 20) ở mức 5%, ta được c = 3,49.

Vì F > c nên ta bác bỏ H_o. Vậy ta khẳng định có tồn tại mối tương quan phi tuyến của Y đối với X. xác suất sai lầm của khẳng định này là 5%.

§ 4. PHÂN TÍCH HỒI QUY TUYẾN TÍNH

Giả sử X là một biến nào đó (có thể là biến ngẫu nhiên hay không ngẫu nhiên), còn Y là một ĐLNN phụ thuộc vào X theo cách sau đây. Nếu X nhận giá trị x, X = x, thì Y sẽ có kỳ vọng là ax + β, ở đó α và β là hằng số và phương sai là σ^2 (không phụ thuộc x). Khi đó ta nói Y có hồi quy tuyến tính theo X, và đường thẳng hồi quy lý thuyết của Y đối với X. các hệ số α, β được gọi là các hệ số hồi quy lý thuyết. X được gọi là biến độc lập, còn Y được gọi là biến phụ thuộc.

Bài toán đặt là hãy ước lượng các hệ số quy lý thuyết α và β trên một mẫu quan (x_1, y_1), ..., (x_n, y_n). Ước lượng α và β dựa trên phương pháp bình phương bé nhất. a và b sẽ được chọn làm ước lượng cho α và β nếu nó làm cực tiểu tổng sau đây:

$$Q(A,B) = \sum_{i=1}^{n} \left(y_i - Ax_i - B\right)^2$$

Hệ phương trình để tìm điểm dừng (a, b) của hàm Q(A, B) có dạng:

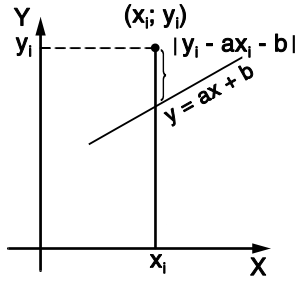
$$\begin{cases} \dfrac{\partial Q}{\partial A} = -2\sum_{i=1}^{n} x_i(y_i - Ax_i - B) = 0 \\ \dfrac{\partial Q}{\partial B} = -2\sum_{i=1}^{n} (y_i - Ax_i - B) = 0 \end{cases}$$

Giải hệ này (hệ phương trình tuyến tính với hai ẩn số A, B), ta tìm được

$$a = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \overline{y} - \overline{ax} = \frac{\sum y - a\sum x}{n}$$

a và b được gọi là các hệ số hồi quy. Đường thẳng với phương trình y = ax + b gọi là đường thẳng hồi quy. Từ cách xác định a, b, ta thấy trong số tất cả các đường thẳng y = Ax + B xuyên qua đám mây điểm $\{(x_i, y_i)\}_{i=1}^{n}$, đường thẳng y = ax + b có tổng bình phương các khoảng cách từ (x_i, y_i) tới đường thẳng là bé nhất.



Ví dụ 11. Các số liệu về số trang của một cuốn sách (X) và giá bán của nó (Y) được cho trong bảng dưới đây.

Tên sách	X	Y (nghìn)
A	400	44
B	600	47
C	500	48
D	600	48
E	400	43

F	500	46

Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.

Giải. Ta có:

$$\sum xy = 138\ 800;$$
$$\sum x = 3000;$$
$$\sum y = 276;$$
$$\sum x^2 = 1540\ 000;$$
$$\sum y^2 = 12718$$

Từ đó

$$a = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$
$$= \frac{6(138800) - (3000)(276)}{6(1540000) - (3000)^2}$$
$$= \frac{4800}{240000} = 0,02;$$
$$b = \frac{\sum y - a\sum x}{n} = \frac{276 - (0,02)(3000)}{6} = 36$$

Vậy đường thẳng hồi quy là

$$y = 0,02x + 36$$ ■

Ngoài việc ước lượng hệ số hồi quy α và β, ta còn quan tâm tới ước lượng σ^2, σ^2 là một con số đo sự phân tán của Y xung quanh đường thẳng hồi quy. Ước lượng cho σ^2, ký hiệu bởi $s^2_{Y.X}$, được cho theo công thức sau:

$$s^2_{Y.X} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b_i)^2$$

Dạng khác của công thức trên là

$$s^2_{Y.X} = \frac{\sum y^2 - a\sum xy - b\sum y}{n-2}$$

Công thức này thường thuận tiện hơn trên thực hành.

$S_{Y.X}$ được gọi là *sai số tiêu chuẩn* của đường hồi quy. Nó cho ta số đo sự phân tán của đám mây điểm (x_i, y_i) xung quanh đường thẳng hồi quy.

Ví dụ 12. Hãy tính sai số tiêu chuẩn của đường hồi quy $S_{Y.X}$ trong ví dụ 11 vừa nêu.

Giải

$$s^2_{Y.X} = \frac{\sum y^2 - a\sum xy - b\sum y}{n-2}$$

$$= \frac{12718 - (0,02).(138800) - 36.(276)}{6-2} = 1,5$$

Vậy: $s_{Y.X} = \sqrt{1,5} = 1,22$

Bây giờ dựa trên phương trình đường thẳng hồi quy tìm được, ta có thể dự báo được giá trị của Y nếu biết giá trị của X. Giá trị được dự báo của Y khi X = x_o sẽ là

$$y_o = ax_o + b$$

Đây đồng thời cũng là giá trị được dự báo cho kỳ vọng của Y ứng với X = x_o (ký hiệu là μ_{x_o}): $\mu_{x_o} = ax_o + b$

Tiếp theo ta xét bài toán tìm khoảng tin cậy cho gia trị dự báo của Y, cũng như khoảng tin cậy cho giá trị dự báo của μ_{x_o}.

+ Công thức để tìm khoảng tin cậy cho giá trị dự báo của Y khi X = x_o sẽ là

$$y_o \pm ts_{X.Y}\sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

trong đó t là phân vị mức $\alpha = \frac{1-\beta}{2}$ của phân bố Student với n – 2 bậc tự do.

+ Công thức để tìm khoảng tin cậy với độ tin cậy β cho giá trị dự báo của μ_{x_o} sẽ là

$$y_o \pm ts_{X.Y}\sqrt{\frac{1}{n}+\frac{(x_o-\bar{x})^2}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

Ví dụ 13. Trở lại ví dụ 11 ta muốn dự báo về giá bán của một cuốn sách với 450 trang.

Giải:

Giá cuốn sách đó được dự báo là

$$y = 0,02.(450) + 36 = 45\,(\text{nghìn})$$

Khoảng tin cậy 95% cho giá của một cuốn sách 450 trang sẽ là

$$45 \pm t.(1,22)\sqrt{1+\frac{1}{6}+\frac{(450-500)^2}{154000-\frac{(3000)^2}{6}}}$$

ở đó t là phân vị mức $\frac{1-0,95}{2} = 0,025$ của phân bố Student với 6 – 2 = 4 bậc tự do. Tra bảng ta tìm được

$$t = 2,776$$

Thay vào công thức trên ta được khoảng tin cậy cần tìm là $45 \pm 3,77$

hay $41,23 < y_o < 48,77$

Vậy với độ tin cậy 95%, cuốn sách với 450 trang sẽ được bán với giá trong khoảng từ 41230 đồng đến 48770 đồng. ■

Ví dụ 14. Trở lại ví dụ 13 ta muốn dự báo giá bán trung bình của tất cả các cuốn sách 450 trang.

Giải. Giá trung bình được sự báo là

$$\mu = 0,02.(450) + 36 = 45$$

Khoảng tin cậy 95% cho giá trung bình của tất cả các cuốn sách 450 trang là

$$45 \pm (2,776).(1,22)\sqrt{\frac{1}{6} + \frac{(450-500)^2}{1540000 - \frac{(3000)^2}{6}}} = 45 \pm 3,4\sqrt{0,23}$$

$$= 45 \pm 1,63$$

hay $43,37 < \mu < 46,63$

Vậy với độ tin cậy 95% giá trung bình của tất cả các cuốn sách 450 trang sẽ nằm trong khoảng từ 43370 đồng đến 46630 đồng. ■

Một vấn đề quan trọng chúng ta phải lưu ý đến là kiểm tra xem hệ số hồi quy lý thuyết α có khác không hay không. Nếu $\alpha = 0$ thì $EY = \beta$ là một hằng số không phụ thuộc X. Khi đó việc dự báo EY dựa trên vô nghĩa. Người ta đã chứng minh được rằng hệ số hồi quy a có độ lệch tiêu chuẩn là

$$s_a = \frac{s_{Y.X}}{s_X\sqrt{n-1}} = \frac{s_{Y.X}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Thống kê: $T = \frac{a}{s_a}$

sẽ có phân bố Student với n – 2 bậc tự do nếu giả thiết H_o: $\alpha = 0$ là đúng. Vì vậy giả thiết H_o sẽ bị bác bỏ ở mức ý nghĩa α nếu $|T| > c$, ở đó c là phân vị mức $\frac{\alpha}{2}$ của phân bố Student với n – 2 bậc tự do.

Ví dụ 15. Với mức ý nghĩa $\alpha = 5\%$, hãy kiểm định giả thiết.

H_o: "Hệ số góc α của đường thẳng hồi quy lý thuyết của Y đối với X bằng không", ở đó X và Y là hai biến xét trong ví dụ 11.

Giải:

Ta có $$s_a = \frac{s_{Y.X}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

$$= \frac{1,22}{\sqrt{1540000 - \frac{(3000)^2}{6}}} = \frac{1,225}{200} = 0,0061$$

Vậy: $T = \dfrac{0,02}{0,006} = 3,33$

Với mức ý nghĩa $\alpha = 5\%$, tra bảng phân bố Student với 4 bậc tự do, ta tìm được $c = t_{0,025} = 2,776$.

Ta có $|T| > c$, do đó ta bác bỏ H_o.

Vậy hệ số góc α của đường thẳng hồi quy lý thuyết của Y đối với X là khác không. ■

Chú thích về sử dụng Minitab

Ta nhập các số liệu của biến độc lập (x_i) vào cột C1 và các số liệu của biến phụ thuộc (y_i) vào cột C2. Sau đó ta gõ lệnh

REGRESS C2 1 C1

Minitab sẽ cho ta trên màn hình phương trình đường thẳng hồi quy mẫu và một bảng phân bố phương sai của bài toán hồi quy. Bảng đó có dạng sau:

Nguồn	**Bậc tự do** (DF)	**Tổng bình phương** (SS)	**Trung bình bình phương** (MS)	**Tỷ số F**
Hồi quy	1	SSR	MSR	$F = \dfrac{MSR}{MSE}$
Sai số	n – 2	SSE	MSE	
Tổng cộng	n – 1	SST		

Ở đây SST là tổng bình phương chung

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

SSR là tổng bình phương do hồi quy

$$SSR = \sum_{i=1}^{n} (ax_i + b - \bar{y})^2$$

còn SSE là tổng bình phương do sai số

$$SSE = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Ta có: SST = SSR + SSE

Có thể chứng minh được rằng:

$$SSR = a^2\left\{\sum x^2 - \frac{(\sum x)^2}{n}\right\} = a\left\{\sum xy - \frac{(\sum x)(\sum y)}{n}\right\}$$

$$SSE = \sum y^2 - a\sum xy - b\sum y$$

Do đó MSE chính là $s^2_{Y.X}$ và tỷ số F chính là $\frac{a^2}{s_a^2}$.

Tỷ số $\frac{SSR}{SST}$ gọi là *hệ số xác định*. Nó chính bằng bình phương hệ số tương quan r^2

$$r^2 = \frac{SSR}{SST}$$

Việc kiểm định giả thiết H_o: "Hệ số góc α của đường thẳng hồi quy lý thuyết của Y đối với X bằng 0", hay tương đương "không có quan hệ hồi quy lý thuyết của Y đối với X bằng 0", hay tương đương "không có quan hệ hồi quy tuyến tính giữa X và Y" mà ta đã trình bày trước đây (dùng test thống kê T = $\frac{a}{s_a}$), nay có thể thay bằng thống kê F = $\frac{MSR}{MSE}$. Giả thiết H_o bị bác bỏ ở mức ý nghĩa α nếu F > c, ở đó c là phân vị mức α của phân bố Fisher với bậc tự do (1, n – 2).

Chẳng hạn bảng phân tích phương sai của bài toán hồi quy trong ví dụ 11 là

Nguồn	**Bậc tự do** (DF)	**SS**	**MS**	**F**
Hồi quy	1	16	16	F = 10,66
Sai số	4	6	1,5	$r^2 = \frac{16}{22} = 0,7272$
Tổng	5	22		r = 0,8528

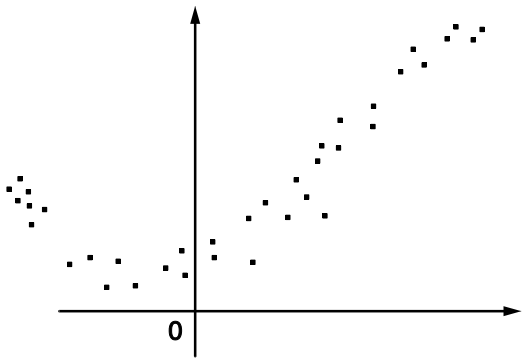
Với mức ý nghĩa α = 5%, tra bảng phân bố Fisher với bậc tự do (1,4) ta được c = 7,71. Vì F = 10,66 > 7,71 nên H_o bị bác bỏ.

'5. HỒI QUY PHI TUYẾN

Nếu khi biến độc lập X nhận giá trị x, biến phụ thuộc Y có kỳ vọng là $\varphi(x)$, ở đó φ là một hàm số nào đó, thì ta gọi $\varphi(x)$ là *hàm hồi quy lý thuyết* của Y đối với X. Trong thực tế có nhiều khi $\varphi(x)$ không phải là một hàm tuyến tính mà có dạng một đa thức bậc 2, bậc 3, ... hay hàm log, sin... Khi đó ta nói *Y có hồi quy*

phi tuyến đối với X. Việc kiểm định xem có hồi quy phi tuyến hay không chúng ta đã trình bày ở mục '3.

Bài toán đặt ra tiếp theo là hãy "ước lượng" hàm hồi quy φ(x) căn cứ trên một mẫu số liệu quan sát được. Hàm hồi quy ước lượng $\varphi(\mathbf{x})$ sẽ phải chọn sao cho nó "gần" với đám mây điểm nhất. Chẳng hạn nếu đám mây điểm có dạng như sau:



ta có thể dự đoán rằng hàm hồi quy φ(x) có dạng một parabol

$$\varphi(x) = Ax^2 + Bx + C$$

Ta sẽ dùng phương pháp bình phương bé nhất để ước lượng các hằng số A, B, C.

Một phương pháp khác cũng hay được áp dụng là phương pháp tuyến tính hóa; giả sử hàm hồi quy lý thuyết có dạng

$$\varphi(x) = Ax^m + B$$

Đặt $Z = x^m$, ta sẽ có hồi quy tuyến tính của Y đối với Z. Dựa trên số liệu $\{x_1, y_1), ..., (x_n, y_n)\}$ ta biến đổi thành số liệu

$$\left\{(\mathbf{x}_1^m, \mathbf{y}_1), ..., (\mathbf{x}_n^m, \mathbf{y}_n)\right\} = \left\{(\mathbf{z}_1, \mathbf{y}_1), ..., (\mathbf{z}_n, \mathbf{y}_n)\right\}$$

ta sẽ ước lượng các hằng số A, B. Theo công thức hệ số hồi quy tuyến tính.

Ví dụ 16. Giả sử hàm hồi quy lý thuyết của Y theo X có dạng sau

$$\varphi(x) = Ax^2 + B$$

Hãy ước lượng φ(x) dựa trên mẫu quan sát sau đây gồm 30 số liệu (x_i, y_i):

x_i	y_i	Tần số	$z_i = \mathbf{x}_i^2$
1	7	4	1

1,5	9,4	4	2,25
2	12,8	2	4
2	13	4	4
2,5	17,6	3	6,25
2,5	17,5	5	6,25
3	23	4	9
3	22,5	2	9
3	22,8	2	9

Giải:

Từ hai cột số liệu (Z, y) ta tìm được $\sum z = 159$; $\sum y = 466,1$; $\sum z^2 = 1080,75$; $\sum y^2 = 8181,83$ và $\sum zy = 2941,27$. ■

Từ đó ước lượng A là a = 2,16, ước lượng của B là b = 3,9. Vậy hàm hồi quy là $y = 2,16x^2 + 3,9$.