PHÂN PHỐI CHI-BÌNH PHƯƠNG
&
XỬ LÝ CÁC TẦN SỐ

Nguyễn Quang Vinh – Nguyễn Thị Từ Vân

CHI-BÌNH PHƯƠNG (χ^2)

Karl Pearson (1857-1936)

- Một trong những phân phối được sử dụng rộng rãi nhất
- Kiểm định giả thuyết khi dữ liệu ở dạng tần số: kiểm sự khác nhau giữa các tỷ lệ
- Phù hợp nhất với các biến số ở dạng phân nhóm / phân loại

N(μ,σ) chuyển thành phân phối bình thường chuẩn N(0,1):

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

$$z^2 = \left(\frac{x - \mu}{\sigma} \right)^2$$ z^2 có phân phối χ^2, với:

độ tự do df = 1, hay:

$$z^2 = \chi^2_{(1)}$$

$$\chi^2_{(2)} = \left(\frac{x_1 - \mu}{\sigma}\right)^2 + \left(\frac{x_2 - \mu}{\sigma}\right)^2 = z_1^2 + z_2^2$$

có phân phối χ^2 với độ tự do df = 2

Nói chung:

$$\chi^2_{(n)} = z_1^2 + z_2^2 + \ldots + z_n^2$$

có phân phối χ^2 với độ tự do df = n

Công thức của phân phối χ^2:

$$f(u) = \frac{1}{\left(\frac{k}{2}-1\right)!}\frac{1}{2^{k/2}}u^{(k/2)-1}e^{-(u/2)}, \quad u > 0$$

với: $e = 2.71828, k = df$

ỨNG DỤNG CỦA χ^2

- Tần số quan sát (Thấy/Thực tế)

  so với

  Tần số kỳ vọng (Nghĩ/Giả thuyết)

- (1) Kiểm định tính phù hợp (mức độ khớp)
  (2) Kiểm định tính độc lập
  (3) Kiểm định tính thuần nhất

χ^2 kiểm định tính phù hợp

- Phép kiểm 2 đuôi cho tỷ lệ
- 2 biến cố:

$$H_O: p = p_0$$
$$H_A: p \neq p_0$$

- Nhiều biến cố:

$$H_O: p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$$

H_A: ít nhất có một tỷ lệ p_i *không phù hợp*

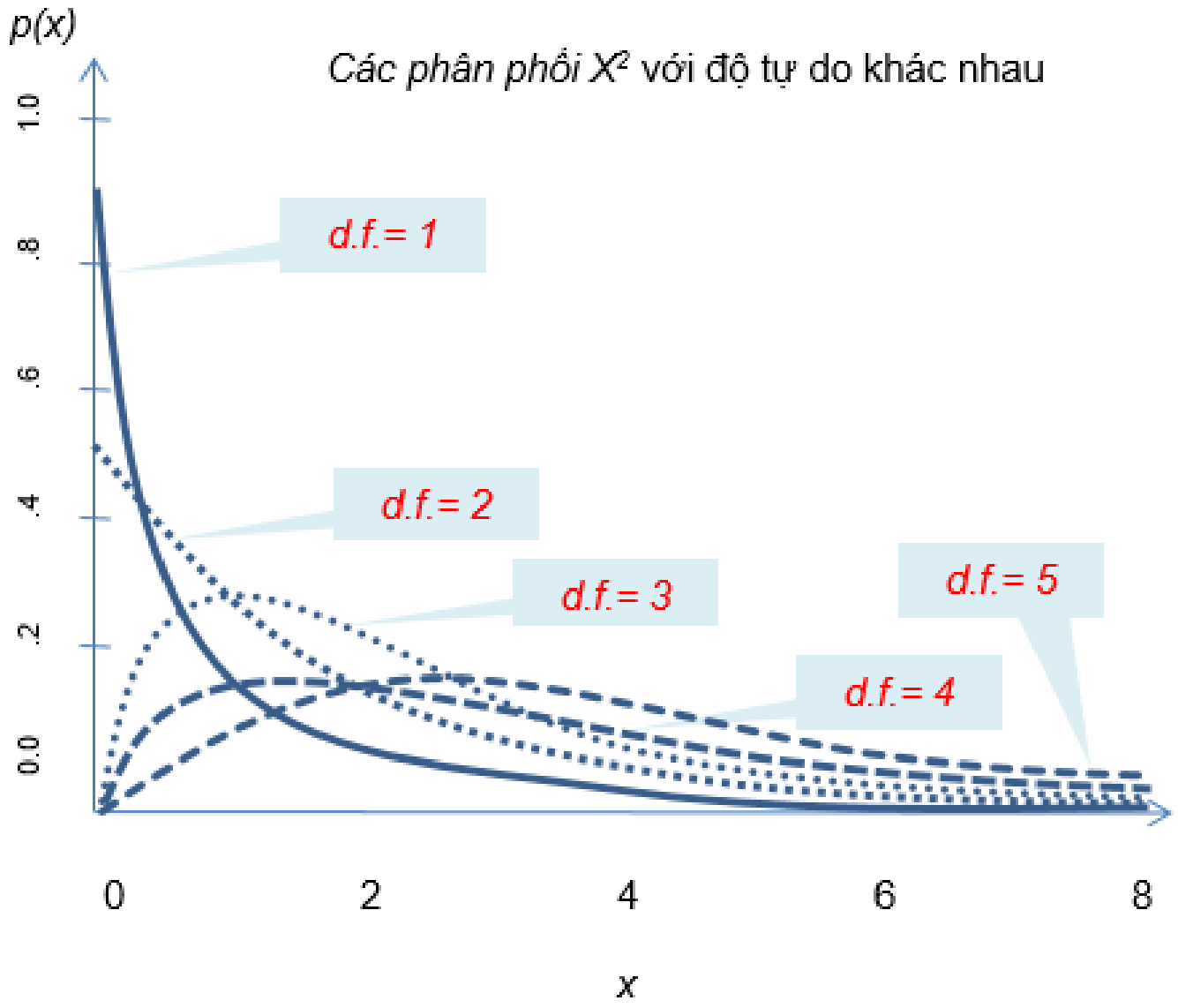
χ^2 kiểm định tính phù hợp

Trị số:

$$\chi_c^2 = \sum \frac{(O-E)^2}{E}$$

Độ tự do = số loại - 1

từ bỏ H_0 *nếu* $\chi_c^2 > \chi^2_{\alpha,df}$


p(x)
Các phân phối X² với độ tự do khác nhau
1.0
.8
.6
.4
.2
0.0
d.f.= 1
d.f.= 2
d.f.= 3
d.f.= 4
d.f.= 5
0
2
4
6
8
x

χ^2 kiểm định tính phù hợp

- *Mức độ phù hợp (khớp) của sự phân bố dữ liệu có được so với một phân phối lý thuyết**
- *Tần số kỳ vọng nhỏ: 5*

  - Kết hợp các nhóm kế cận → để đạt được số tối thiểu.

**Kolmogorov-Smirnov kiểm → các phân phối liên tục*

χ^2 kiểm định tính độc lập

- Phép kiểm χ^2 được dùng nhiều nhất
- *Một* tổng thể, khi mỗi cá thể được phân loại theo 2 tiêu chuẩn:

  Tiêu chuẩn thứ 1: hàng

  Tiêu chuẩn thứ 2: cột

- Bảng phân loại theo 2 tiêu chẩn: r hàng, c cột
- H_O: 2 tiêu chuẩn phân loại *độc lập với nhau (không có liên quan)*

  H_A: 2 tiêu chuẩn phân loại *không độc lập với nhau (có liên quan)*

- $df = (r - 1)(c - 1)$

χ^2 kiểm định tính độc lập

Tần số kỳ vọng nhỏ

- Không nên dùng phép kiểm χ^2 test nếu *có bất kỳ* $E_i < 5$

χ^2 kiểm định tính thuần nhất

Để xác định xem các nhóm riêng biệt có thể được xem là cùng thuộc một tổng thể.

χ^2 *kiểm định tính độc lập*

- Tổng số ở hàng và cột không bị kiểm soát bởi người làm nghiên cứu
- *? có sự liên quan* (2 tiêu chuẩn)

χ^2 *kiểm định tính thuần nhất*

- Tổng số ở hàng hay cột bị kiểm soát bởi người làm nghiên cứu
- *? có đồng nhất* (các mẫu nghiên cứu có phải từ một tổng thể)

cách tính toán ***như nhau*** *nhưng ý niệm* ***khác nhau***

PHÉP KIỂM CHÍNH XÁC FISHER

	Điều trị	Chứng	**Tổng**
O+	x	$K - x$	**K**
O-	$n - x$	$(N-K)-(n-x)$	**N - K**
Tổng	**n**	**N-n**	**N**

$$N \rightarrow \begin{cases} K & x \\ N-K & n-x \end{cases} \leftarrow n$$

$$P(x) = \frac{{}_{K}C_{x} \cdot {}_{N-K}C_{n-x}}{{}_{N}C_{n}}$$

Chúng ta có kết quả của thực nghiệm như sau:

	Điều trị	Chứng	**Tổng**
O+	6	1	**7**
O-	2	4	**6**
Tổng	**8**	**5**	**13**

Liệt kê tất cả các tình huống có thể có trong một cỡ mẫu 13, có được:

- 7 kết cục tốt &
- 8 đối tượng trong nhóm điều trị.

Chúng ta có 6 bảng sau:

	Điều trị	Chứng	**Tổng**
O+	***7***	0	**7**
O-	1	5	**6**
Tổng	**8**	**5**	**13**

$$P(x=7)=\frac{{}_{7}C_{7}\cdot{}_{6}C_{1}}{{}_{13}C_{8}}$$

$$=\frac{6}{1287}=.0047$$

	Điều trị	Chứng	**Tổng**
O+	***6***	1	**7**
O-	2	4	**6**
Tổng	**8**	**5**	**13**

$$P(x=6)=\frac{{}_{7}C_{6}\cdot{}_{6}C_{2}}{{}_{13}C_{8}}$$

$$=.0816$$

	Điều trị	Chứng	**Tổng**
O+	***5***	2	**7**
O-	3	3	**6**
Tổng	**8**	**5**	**13**

$$P(x=5)=\frac{{}_{7}C_{5}\cdot{}_{6}C_{3}}{{}_{13}C_{8}}=.3262$$

	Điều trị	Chứng	**Tổng**
O+	***4***	3	**7**
O-	4	2	**6**
Tổng	**8**	**5**	**13**

$$P(x=4)=\frac{{}_{7}C_{4}\cdot{}_{6}C_{4}}{{}_{13}C_{8}}=.4070$$

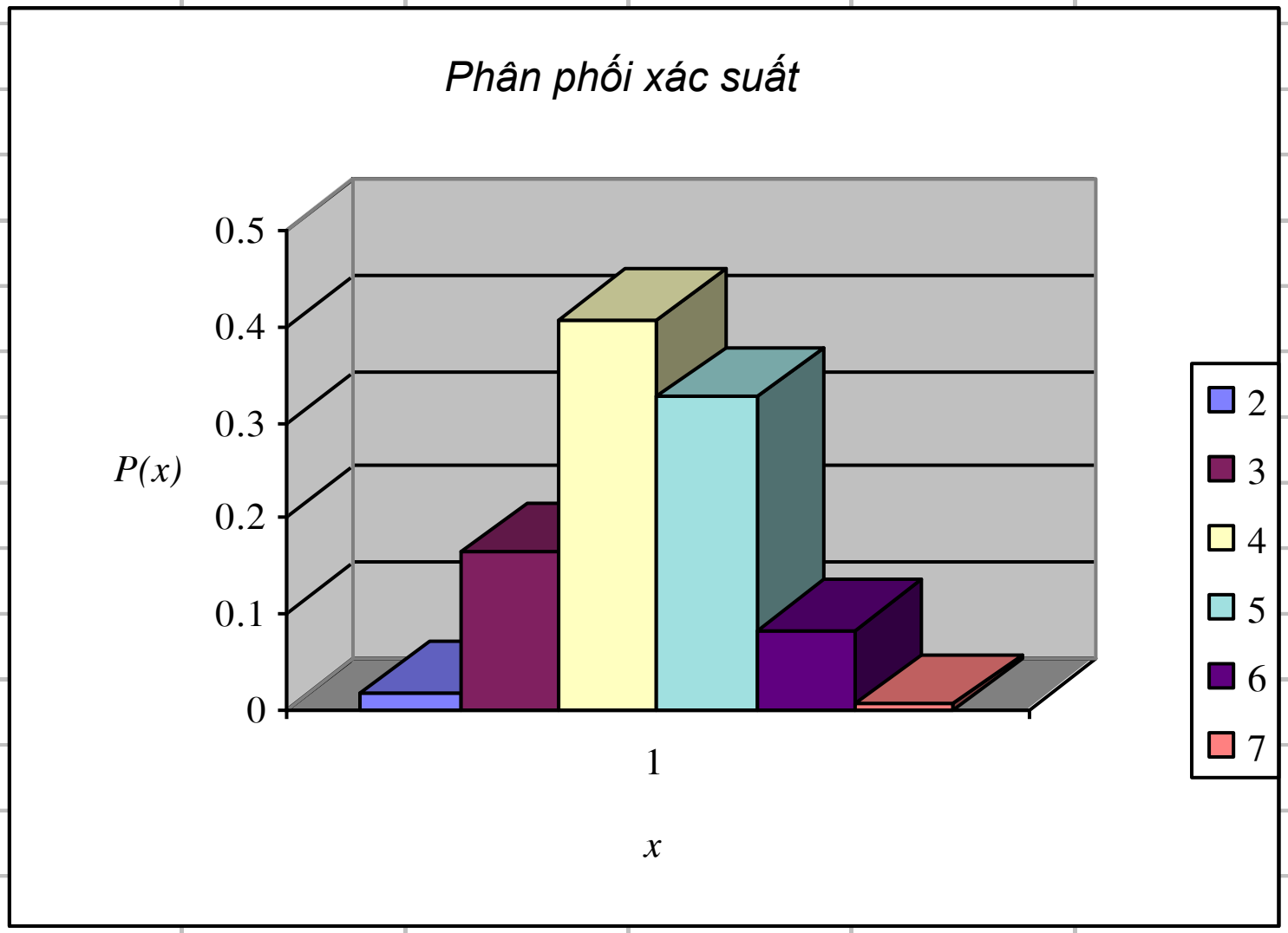
	Điều trị	Chứng	**Tổng**
O+	***3***	4	**7**
O-	5	1	**6**
Tổng	**8**	**5**	**13**

$$P(x=3)=\frac{{}_7C_{3\cdot 6}\,C_5}{{}_{13}C_8}=.1632$$

	Điều trị	Chứng	**Tổng**
O+	***2***	5	**7**
O-	6	0	**6**
Tổng	**8**	**5**	**13**

$$P(x=2)=\frac{{}_7C_{2\cdot 6}\,C_6}{{}_{13}C_8}=.0163$$

***Kiểm lại: cộng tất cả các xác suất = 1 (có làm tròn số)**


Phân phối xác suất
0.5
0.4
0.3
0.2
0.1
0
P(x)
1
x
2
3
4
5
6
7

Giả thuyết

- H_O: $\pi_T = \pi_C$
  (không có sự khác biệt giữa 2 nhóm: điều trị và chứng)
- H_A: $\pi_T > \pi_C$ *(1-đuôi)*

  hay,

  H_A: $\pi_T \neq \pi_C$ *(2-đuôi)*

Tính giá trị P

- Xác suất để có bộ dữ liệu nghiên cứu là 0.0816
- Giá trị P là xác suất để có bộ dữ liệu, hay hơn nữa (tốt hơn/xấu hơn).

Giá trị P một đuôi:

$P(x \geq 6) = P(x=6) + P(x=7) = 0.0816 + 0.0047 = 0.0863$

Tính giá trị P

Giá trị P hai đuôi:

(1) $P(x \geq 6 \text{ hay } x \leq 2) = P(x=2)+P(x=6)+P(x=7) =$
$0.0816 + 0.0047 + 0.0163 = 0.1026$

(2) Nhân đôi kết quả một đuôi*, thành:

$P = 2x0.0863 = 0.1726$

**tính xấp xỉ*