ĐẠI HỌC Y DƯỢC CẦN THƠ
BỘ MÔN THỐNG KÊ - DÂN SỐ

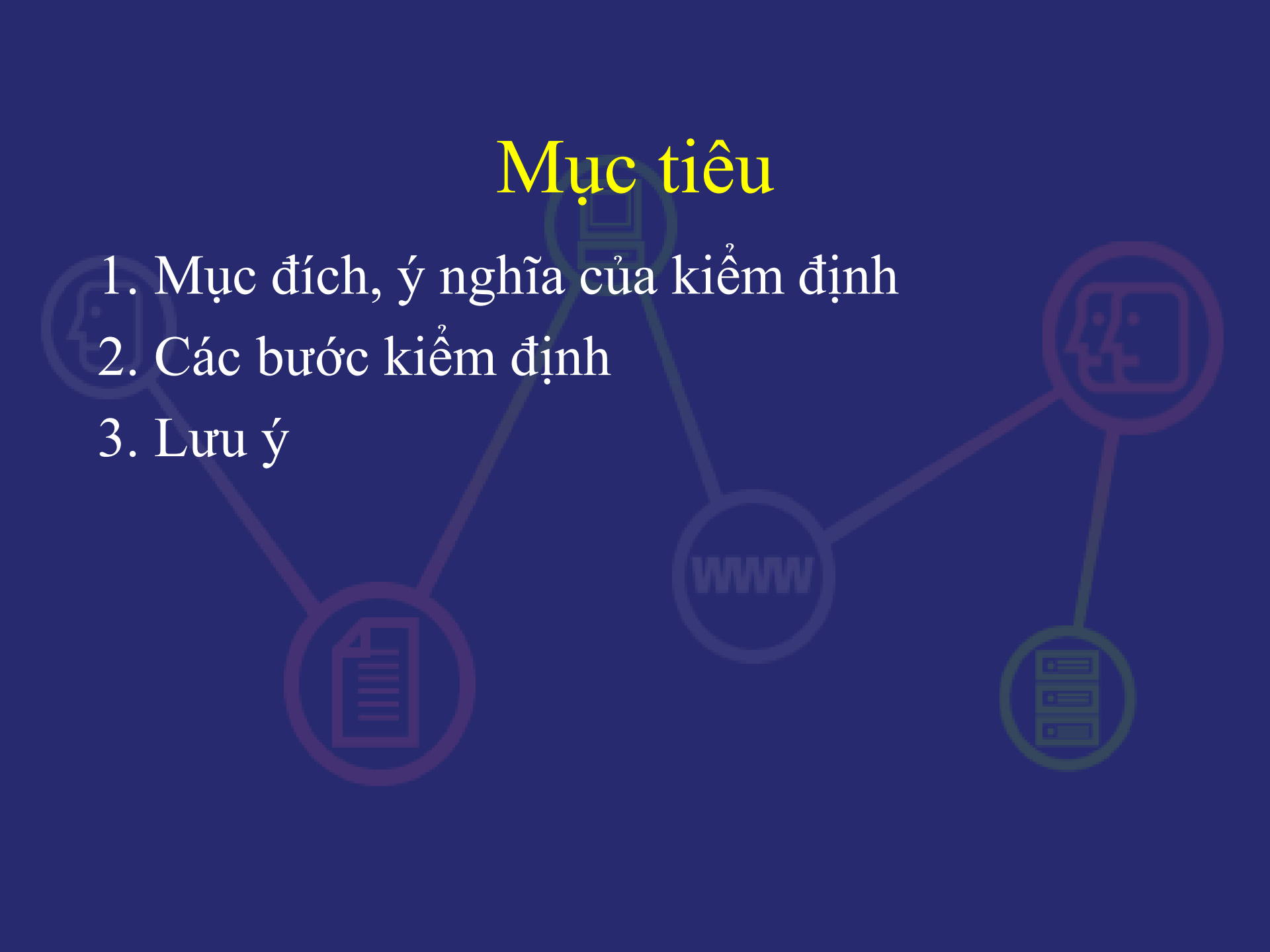
Kiểm định Khi bình phương

χ^2

ThS. Nguyễn Chí Minh Trung

Mục tiêu

1. Mục đích, ý nghĩa của kiểm định
2. Các bước kiểm định
3. Lưu ý

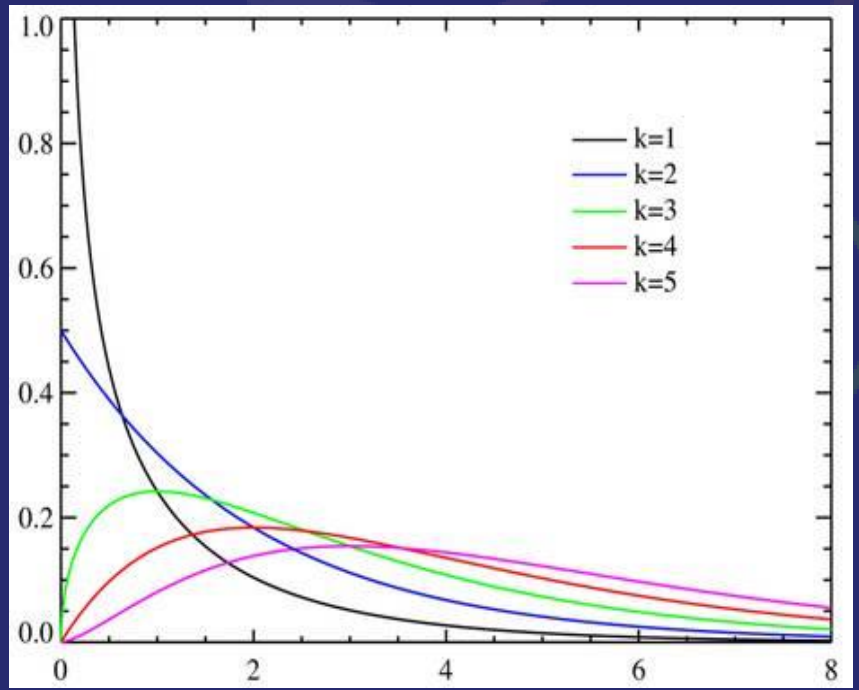


Phân phối Khi bình phương χ2

$$\chi^2 = \sum_{i=1}^{k} Z_i^2$$

có luật phân phối khi phương bậc tự do *k*

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



Bảng tiếp liên

- Bảng phân bố tần số hai chiều còn được gọi là bảng tiếp liên
- Số hàng và số cột tương ứng với số phân nhóm (mức độ) của hai biến
- Con số trong bảng là số người thể hiện sự kết hợp các mức độ tương ứng của hai biến

Ứng dụng của kiểm định χ^2

- Kiểm định χ^2 dùng trong nhiều trường hợp:

1. Kiểm định tính phù hợp (*goodness-of-fit*),
2. **Kiểm định tính độc lập (*independence*),**
3. Kiểm định tính đồng nhất (*homogeneity*).

- Kiểm định χ^2 cũng dùng để so sánh hai tỷ lệ
- Kiểm định χ^2 Mantel-Haenszel để hiệu chỉnh yếu tố nhiễu

Phân tích bảng tiếp liên

- Bảng tiếp liên thể hiện mối quan hệ giữa hai biến phân loại.
- Độc lập: phân bố của một biến giống nhau giữa tất cả các mức độ của biến kia
- Không độc lập (liên quan): phân bố của một biến không giống nhau giữa các mức độ của biến kia

Hai biến *tiêm vắc xin* và *mắc cúm* độc lập hay liên quan với nhau ???

	Vắc xin	**Placebo**	**Tổng**
Cúm	20 (8,3%)	80 (36,4%)	**100**
Không cúm	220	140	**360**
Tổng	**240**	**220**	**460**

Tần số kỳ vọng

Nếu không có mối liên quan giữa việc ***tiêm vắc xin*** và việc ***mắc cúm***, thì tần số kỳ vọng sẽ bằng

	Vắc xin	**Placebo**	**Tổng**
Cúm	a	b	**a + b (100)**
Không cúm	c	d	**c + d (360)**
Tổng	**a + c (240)**	**b + d (220)**	**n (460)**

Tần số kỳ vọng

- Tỷ lệ mắc cúm trong nhóm tiêm và không tiêm vắc xin là như nhau, và bằng tỷ lệ mắc cúm chung. $$\frac{a}{a+c} = \frac{b}{b+d} = \frac{a+b}{n}$$
- Do đó tần số kỳ vọng $$a = \frac{(a+c)\times(a+b)}{n}$$
- Tần số kỳ vọng $$= \frac{\sum \text{hang} \times \sum \text{cot}}{\sum chung}$$

Tính tần số kỳ vọng

Nếu không có mối liên quan giữa việc ***tiêm vắc xin*** và việc ***mắc cúm***, thì tần số kỳ vọng sẽ bằng

	Vắc xin	Placebo	Tổng
Cúm	52,2	47,8	100
Không cúm	187,8	172,2	360
Tổng	240	220	460

	Vắc xin	Placebo	Tổng
Cúm	20	80	100
Không cúm	220	140	360
Tổng	240	220	460

Tần số quan sát (Observed)

Tần số kỳ vọng (Expected)

	Vắc xin	Placebo	Tổng
Cúm	52,2	47,8	100
Không cúm	187,8	172,2	360
Tổng	240	220	460

- So sánh sự khác biệt giữa tần số quan sát (O) với tần số kỳ vọng (E)
  - Tần số quan sát (*O*): tần số thực sự thu được từ mẫu ngẫu nhiên
  - Tần số kỳ vọng (*E*): tần số dự đoán khi giả định hai biến độc lập nhau
- $\sum \frac{(O-E)^2}{E}$ tuân theo phân bố khi bình phương với (r-1)(c-1) bậc tự do
  - r là số hàng và c là số cột

- χ^2 có phân bố dương
- χ^2 chỉ bằng 0 khi tần số quan sát bằng tần số kỳ vọng (O = E)
- Sự khác biệt giữa O và E càng lớn, thì
  - ✓ giá trị χ^2 càng lớn
  - ✓ Sự khác biệt đó (mối liên quan giữa hai biến) càng ít khả năng là do ngẫu nhiên

Kiểm định khi bình phương

- Để xác định mối liên quan, dùng kiểm định khi bình phương của Pearson
- Còn gọi là kiểm định tính độc lập
- Khi kiểm định, nhà nghiên cứu thường mong muốn:
  - chứng minh có mối liên quan giữa hai biến (giả thuyết H_1), và
  - bác bỏ tính độc lập (giả thuyết H_0)

Các bước tiến hành kiểm định

- Mô tả số liệu
- Giả định: mẫu nhẫu nhiên
- Giả thuyết: H_0=độc lập/H_1=không độc lập
- Kiểm định:$\chi^2 = \sum \frac{(O-E)^2}{E}$

Phân bố xác suất: phân bố xác suất xấp xỉ phân phối khi bình phương với df:(r-1)x(c-1)

- Mức ý nghĩa:

0,05->3,84; 0,01->6,63; 0,001->10,83

- Tính
- Kết luận:χ2 tính được>χ2 tra bảng bỏ H_0

Các bước tiến hành kiểm định

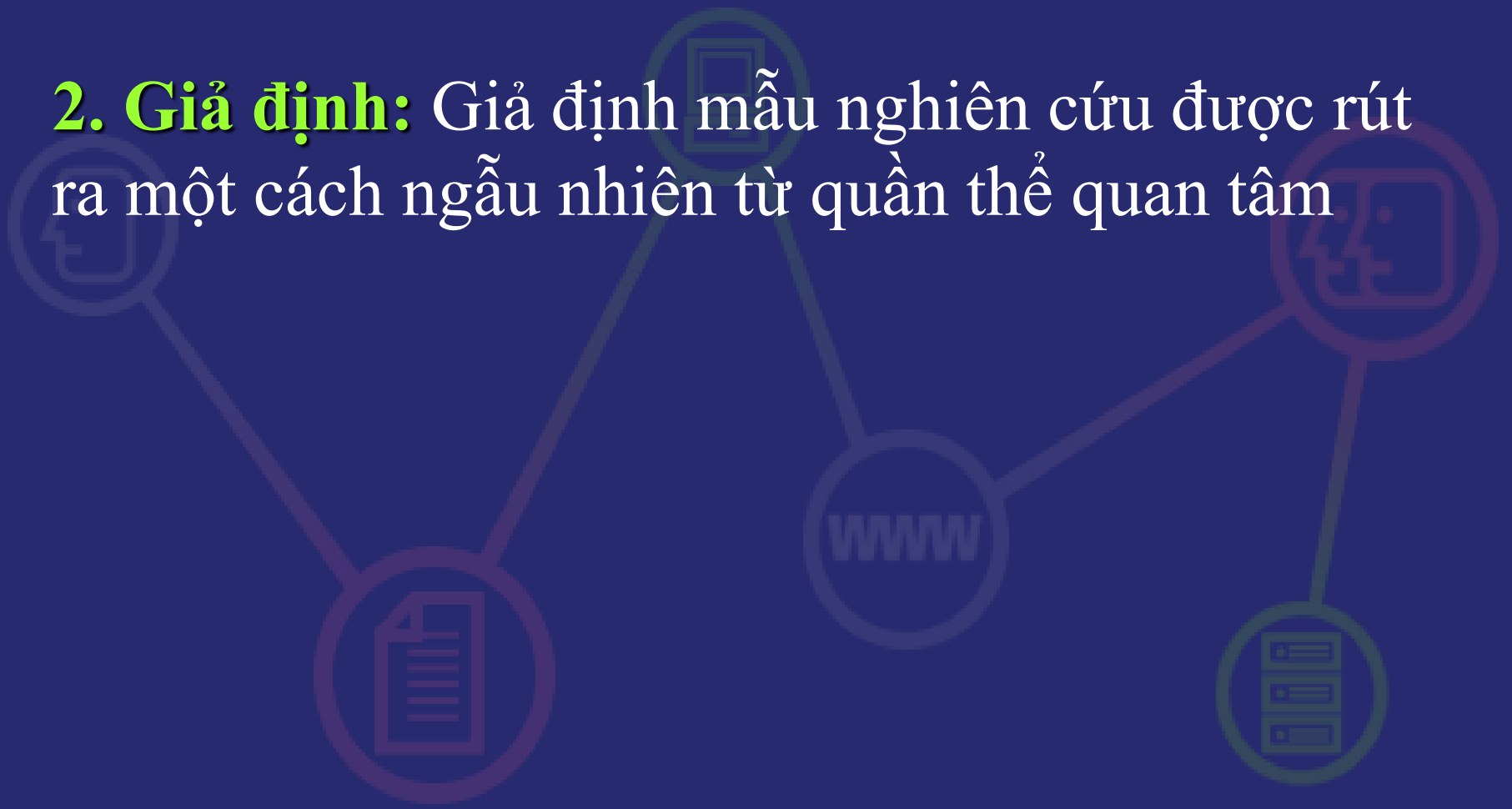
1. Mô tả số liệu

Cúm	Vắc xin	Placebo	Tổng
Có	20 (8.3%)	80 (36,4%)	100 (21.7%)
Không	220	140	360
Tổng	240	220	460

Chúng ta cần tìm hiểu xem tiêm vắc xin có làm giảm nguy cơ mắc cúm không?

Các bước tiến hành kiểm định

2. Giả định: Giả định mẫu nghiên cứu được rút ra một cách ngẫu nhiên từ quần thể quan tâm



Các bước tiến hành kiểm định

3. Giả thuyết/Đối giả thuyết

H_0: Hai biến *mắc cúm* và *loại thuốc dùng* (vắc xin hay placebo) là độc lập với nhau.

H_1: Hai biến trên không độc lập (hay có mối liên quan với nhau).

Các bước tiến hành kiểm định

4. Thống kê để kiểm định và phân phối xác suất

•Kiểm định: $\chi^2 = \sum \frac{(O-E)^2}{E}$

Có phân bố xác suất xấp xỉ phân phối khi bình phương với df:(2-1)x(2-1)=1

Trong đó:

- O: các tần số quan sát được (*observed*) trên thực tế
- E: các tần số kỳ vọng *(Expected)* khi không có mối liên quan giữa hai biến nói trên.

Các bước tiến hành kiểm định

5. Chọn mức ý nghĩa thích hợp

Với 1 bậc tự do:

- $\alpha = 0{,}05$ => *giá trị tra bảng* $\chi^2 = 3{,}84$
- $\alpha = 0{,}01$ => *giá trị tra bảng* $\chi^2 = 6{,}635$
- $\alpha = 0{,}001$ => *giá trị tra bảng* $\chi^2 = 10{,}83$

=> Bác bỏ H_0 nếu giá trị χ^2 tính được $\geq$ giá trị χ^2 tra bảng

Các bước tiến hành kiểm định

6. Tính toán cụ thể $E = \frac{\sum \text{hang} \times \sum \text{cot}}{\sum chung}$

	Vắc xin	Placebo	Tổng
Cúm	a	b	100
Không cúm	c	d	360
Tổng	240	220	460

Ví dụ: tần số kỳ vọng a = (100 × 240)/460 = 52,2

Các bước tiến hành kiểm định

6. Tính toán cụ thể

$$E = \frac{\sum \text{hang} \times \sum \text{cot}}{\sum chung}$$

	Vắc xin	*Placebo*	*Tổng*
Cúm	20 52,2	80 47,8	100
Không cúm	220 187,8	140 172,2	360
Tổng	240	220	460

Các bước tiến hành kiểm định

6. Tính toán cụ thể

$$\chi^2_{kd} = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(20-52,2)^2}{52,2} + \frac{(80-47,8)^2}{47,8} + \frac{(220-187,8)^2}{187,8} + \frac{(140-172,2)^2}{172,2}$$

$$= 19,86 + 21,69 + 5,52 + 6,02 = 53,09$$

Các bước tiến hành kiểm định

7. Kết luận kiểm định

53,09 > 10,83 (giá trị χ^2 tra bảng với một bậc tự do ở mức ý nghĩa $\alpha = 0{,}001$)

➔*Bác bỏ H_0, chấp nhận H_1 ở mức ý nghĩa α=0,001*

Có mối liên quan giữa hai biến *tiêm vắc xin* và *mắc bệnh cúm*, $\chi_1^2 = 53{,}09$, n=460 p<0,001

Vì tỷ lệ mắc cúm ở nhóm dùng vắc xin (8,3%) nhỏ hơn nhóm dùng placebo (36,4%), có thể kết luận vắc xin thực sự có hiệu quả

Kiểm định χ^2 với bảng 2x2

- Có thể hiệu chỉnh chính xác hơn bằng hiệu chỉnh liên tục của Yates

$$\chi^2_{kd} = \sum \frac{(|O-E| - \frac{1}{2})^2}{E}$$

$$= \frac{\left(32,2 - \frac{1}{2}\right)^2}{52,2} + \frac{\left(32,2 - \frac{1}{2}\right)^2}{47,8} + \frac{\left(32,2 - \frac{1}{2}\right)^2}{187,8} + \frac{\left(32,2 - \frac{1}{2}\right)^2}{172,2}$$

$$= 19,25 + 21,02 + 5,35 + 5,84 = 51,46, \quad p < 0,001$$

Kiểm định χ^2 với bảng 2x2

Cách tính nhanh:

- Ký hiệu giá trị thực của các ô trong bảng

	Vắc xin	**Placebo**	**Tổng**
Cúm	a	b	**e**
Không cúm	c	d	**f**
Tổng	**g**	**h**	**n**

Kiểm định χ^2 với bảng 2x2

Cách tính nhanh:

$$\chi^2_{kd} = \frac{n(ad-bc)^2}{efgh}$$

$$= \frac{460 \times (20 \times 140 - 80 \times 220)^2}{100 \times 360 \times 240 \times 220} = 53,01$$

Kiểm định χ^2 với bảng 2x2

Cách tính nhanh với hiệu chỉnh liên tục:

$$\chi^2_{kd} = \frac{n\left(|ad - bc| - n/2\right)^2}{efgh}$$

$$= \frac{460 \times (14800 - 230)^2}{100 \times 360 \times 240 \times 220} = 51,37$$

So sánh với kiểm định chuẩn Z

- Kiểm định χ^2 cho bảng 2x2 tương đương với kiểm định z so sánh hai tỷ lệ: $\chi^2=z^2$

So sánh với kiểm định chuẩn Z

	Vắc xin	Placebo	Tổng
Cúm	20 (r_1)	80 (r_2)	**100**
Không cúm	220	140	**360**
Tổng	**240 (n_1)**	**220 (n_2)**	**460**

So sánh với kiểm định chuẩn Z

$$z = \frac{p_1 - p_2}{se_{p_1 - p_2}} = \frac{\left(\frac{r_1}{n_1} - \frac{r_2}{n_2}\right)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p = \frac{r_1 + r_2}{n_1 + n_2} = \frac{20 + 80}{240 + 220} = 0,22$$

So sánh với kiểm định chuẩn Z

$$z = \frac{\left(\frac{20}{240} - \frac{80}{220}\right)}{\sqrt{0,22(1-0,22)\left(1/240 + 1/220\right)}} = 7,29$$

$$z^2 = 7,29^2 = 53,08 = \chi^2$$

So sánh kiểm định χ^2 và kiểm định Z

- Từ kiểm định z có thể tính được khoảng tin cậy
- Kiểm định χ^2 dễ áp dụng
- Có thể mở rộng để so sánh nhiều tỷ lệ

Tóm tắt

1. Kiểm định χ^2 dùng để kiểm định mối quan hệ giữa hai biến phân loại
2. Có liên quan tới kiểm định chuẩn
3. Bảng 2x2: khi các ô trong bảng quá nhỏ:
   - Tổng chung của bảng n < 20
   - 20 < tổng chung < 40 và tần số dự tính nhỏ nhất < 5

   $\Rightarrow$ áp dụng kiểm định chính xác của fisher
4. Bảng lớn: dưới 1/5 số ô có tần số dự tính <5 và không có giá trị nào < 1

Sử dụng SPSS

1. Mô tả một biến phân loại: *Analyze\Descriptive Statistics\Frequencies*

2. Mô tả mối liên quan từ 2 biến trở lên:

Analyze\Descriptive Statistics\Crosstabs: Cells\Row

3. Kiểm định giả thuyết cho 1 tỷ lệ

Analyse → Nonparametric Tests → Legacy Dialogs → Chi-Square.

4. Kiểm định giả thuyết cho 2 hay nhiều tỉ lệ

Analyse \Descriptive statistics\Crosstabs: Statistics\Chi-Square.