Bài giảng 6:
Trị số P, kiểm định thống kê, kiểm định giả thuyết

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

P-value

Bảng 3: Giá trị creatinin trong huyết thanh ở đối tượng nghiên cứu

Nhóm THA	N	Mean	SD	CI – 95%	P so với nhóm không THA
Không THA	86	70,27	15,78	66,88 – 73,65	
THA độ 1	221	71,58	17,75	69,22 – 73,93	>0,05
THA độ 2	134	75,07	22,09	71,29 – 78,85	=0,08
THA độ 3	24	81,37	18,13	73,71 – 89,02	<0,01

Nhận xét: Độ THA càng cao thì giá trị Cre trong huyết thanh càng lớn. **Có sự khác biệt có ý nghĩa thống kê với p<0,01** ở nhóm THA độ 3 so với nhóm không THA. Giá trị Cre trong huyết thanh ở các đối tượng bị THA độ 2 cao hơn nhóm không THA (71,29 – 78,85) so với 66,88 – 73,65 mmol/L) **tuy nhiên sự khác biệt này mới đạt p=0,08**.

Tạp chí Y học 2005, 39(6), trang 43

P-value (breast cancer and fat consumption)

- Women's Health Initiative Study (WHI), JAMA

"*A low fat dietary pattern did not result in a statistically significant reduction in invasive breast cancer risk*"

Data:

Invasive breast cancer HR 0.91 (0.83 – 1.01), P = 0.07

Breast cancer mortality HR 0.77 (0.48 – 1.22)

Cancer risks

- Electric razors
- Broken arms (women)
- Fluorescent lights
- Allergies
- Breeding reindeer

- **Being a waiter**
- **Owning a pet bird**
- **Being short**
- **Being tall**
- **Hot dogs**
- **Have a refrigerator!**

Altman and Simon1992, JNCI

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Yếu tố ngẫu nhiên!

About 25% of all findings with "p<0.05" should, if viewed in a scientifically agnostic light, properly be regarded as nothing more than chance findings

Khoảng 25% tất cả phát hiện với P<0.05 nếu đặt dưới lăng kính khoa học khách quan có thể xem là những phát hiện ngẫu nhiên (chứ chẳng có khám phá thật nào cả)

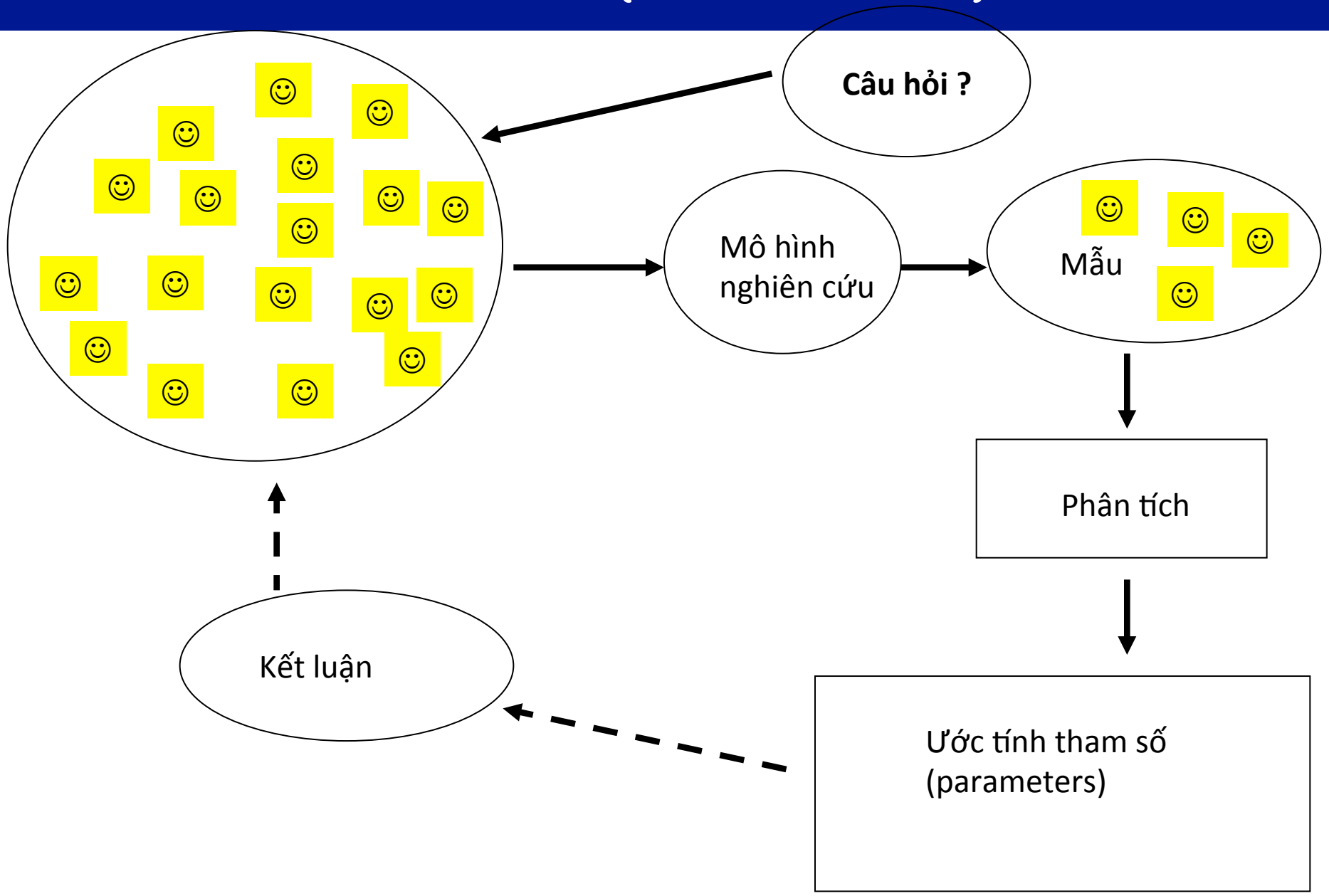
J. Berger (1987); R Matthews (2001)

Ba lĩnh vực trong thống kê học

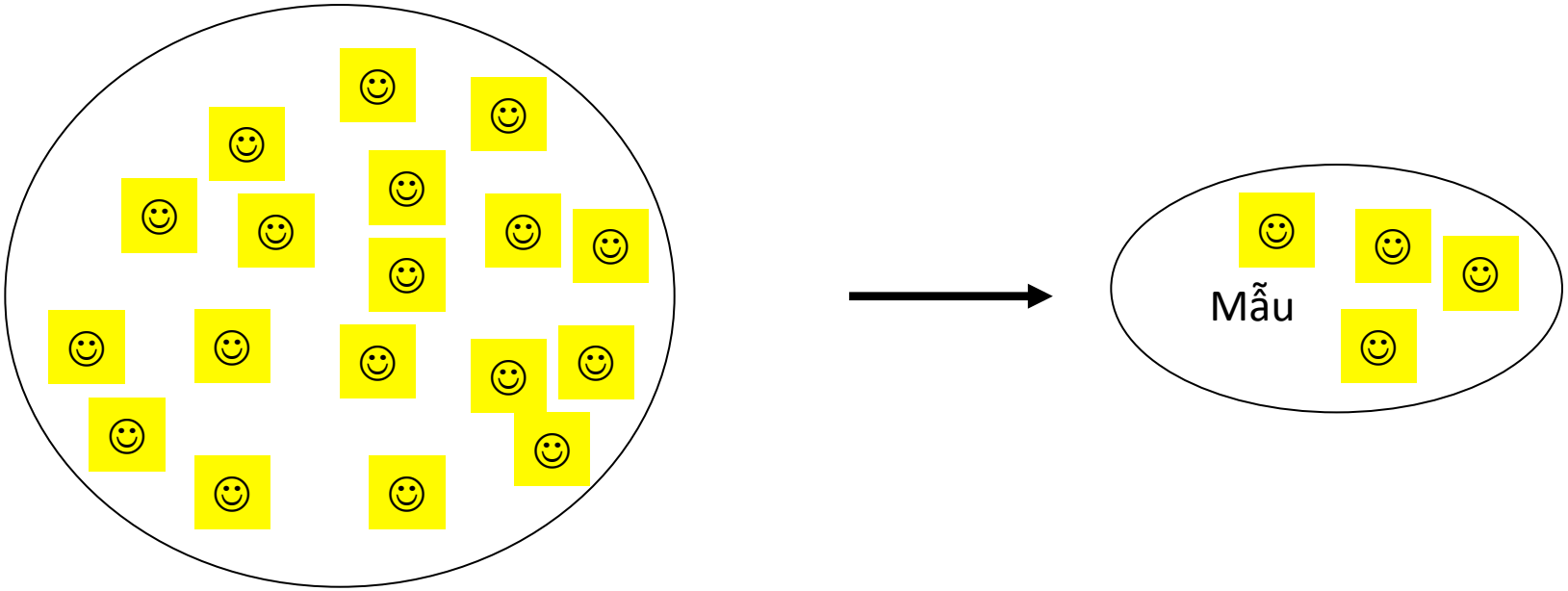
- Ước tính – **Estimation**
- Kiểm định thống kê – **test of significance**
- Kiểm định giả thuyết – **test of hypothesis**

Ước tính (estimation)

Ước tính (estimation)



Ước tính (estimation)



Thông số (parameters)
Trung bình (μ)
Phương sai (σ^2)
Độ lệch chuẩn (σ)
v.v.

Ước số (estimates)
Trung bình (μ)
Phương sai (s^2)
Độ lệch chuẩn (s)
Khoảng tin cậy 95%
v.v.

Không quan tâm đến trị số P và giả thuyết

Ví dụ về estimation – ước tính

- Không biết bao nhiêu người mắc bệnh tiểu đường
- Giải pháp:
  - Chọn mẫu ngẫu nhiên n = 1000 cá nhân.
  - Đo nồng độ đường trong máu; đếm số người mắc bệnh m = 100.
  - Tỉ lệ mắc bệnh 10% (khoảng tin cậy: 8.1 – 11.9%)
- **Estimate = Tỉ lệ mắc bệnh**

Ví dụ về estimation

Women's Health Initiative Study (WHI), JAMA

"A low fat dietary pattern did not result in a statistically significant reduction in invasive breast cancer risk"

Dữ liệu:

Invasive breast cancer HR 0.91 (0.83 – 1.01)

Breast cancer mortality HR 0.77 (0.48 – 1.22)

Test of significance
(kiểm định thống kê)

Kiểm định thống kê

- Đề xuất bởi Ronald Fisher
- Thống kê là một khoa học *qui nạp (inductive inference)*: kết luận đi từ mẫu và áp dụng cho quần thể, từ nhỏ đến lớn.
- Chịu ảnh hưởng bởi **lí thuyết phản nghiệm** (falsificationism) của Karl Popper

Test of significance (Fisher)

- Falsificationism (chủ nghĩa phản nghiệm)
- Phát biểu giả thuyết vô hiệu - **null hypothesis (H0)**
- Thu thập dữ liệu – **data (D)**
- Tính xác suất **data nếu giả thuyết vô hiệu đúng**

P(D | H0)

"A null hypothesis can be disproved, but never proved or established" (Fisher, 1925)

Ngạc nhiên?

Choáng với chiều cao "cực khủng" của nam sinh lớp 12

LinkHay Thích 1.613 người thích nội dung này. Hãy là người đầu tiên trong số bạn bè của bạn.

Lâm Phương - theo Trí Thức Trẻ | 27/04/2013 19:00 Chia sẻ:

ĐỌC NHIỀU NHẤT

(Soha.vn) - Sở hữu chiều cao tới 2,04m, em Hồ Nguyễn Đức Tài (sinh năm 1994) học sinh lớp 12 ngụ tại TP.HCM làm cho nhiều người ngỡ ngàng.

Ngạc nhiên?

Đặng Thu Thảo – Hoa hậu Việt Nam 2012 – Hoa hậu Người Bạc Liêu

629 likes · 5 talking about this

Public Figure

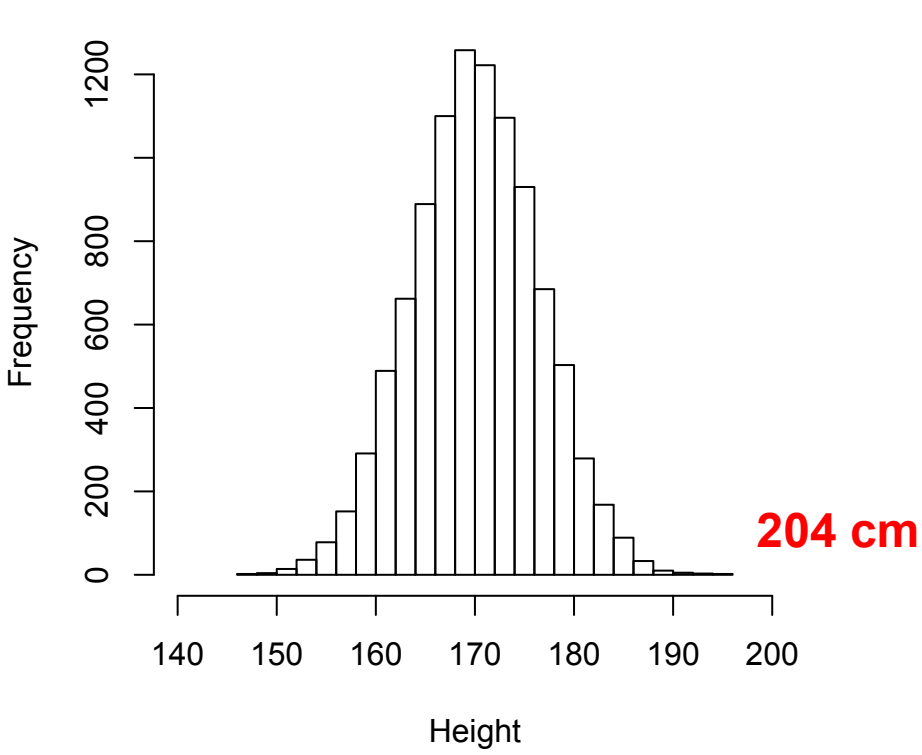
Họ tên: Đặng Thu Thảo Năm sinh: 1991 Chiều cao: 1,73 m Cân nặng: 49 kg Số đo ba vòng 83-60-90

Ngạc nhiên phải đặt trong bối cảnh

- Chiều cao trung bình của thanh niên Việt Nam
  - Nam: 170 cm (SD 6.3 cm)
  - Nữ: 156 cm (SD 6.0 cm)
- Tuân theo luật phô bố chuẩn
- Chiều cao của hai cá nhân so sánh với chiều cao của quần thể thanh niên Việt Nam?

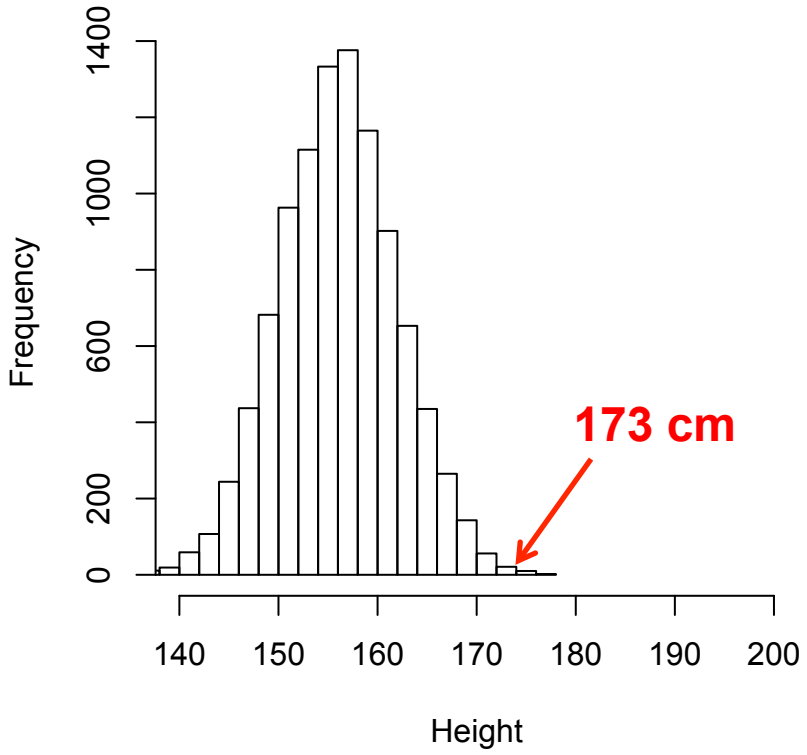
Mô phỏng chiều cao trong cộng đồng

Nam



Height=rnorm(10000, mean=170, sd=6.3)
hist(Height, xlim=c(140,200), breaks=20)

Nữ



Height=rnorm(10000, mean=156, sd=6.0)
hist(Height, xlim=c(140,200), breaks=20)

Định lượng hoá "ngạc nhiên"

- Chiều cao trong quần thể ~ N(170, 6.3)
- Xác suất một thanh niên trong quần thể đó với chiều cao 204 cm?
- Dùng R:

```
1 - pnorm(204, mean=170, sd=6.3)
```

- Trả lời: 0.000000034 (3 trên 100 triệu)

Trị số P

- Trị số P chính là một thước đo về sự ngạc nhiên
- Để tính trị số P chúng ta cần
  - Giá trị tham chiếu (và phân bố)
  - Giá trị thực tế muốn so sánh
- Ý nghĩa của trị số P

$$P(\text{data} \mid \text{reference})$$

Một ví dụ khác về test of significance

10 bệnh nhân được điều trị bằng 2 loại thuốc (A và B). Kết quả trên 8 bệnh nhân cho thấy B>A. **Có thật sự B>A ?**

ID	A	B	B>A
1	1.00	1.02	Yes
2	0.76	0.80	Yes
3	0.89	0.85	No
4	0.70	0.73	Yes
5	0.90	0.92	Yes
6	0.88	0.93	Yes
7	0.92	0.95	Yes
8	0.80	0.82	Yes
9	0.72	0.78	Yes
10	1.10	1.08	No

Qui trình kiểm định thống kê

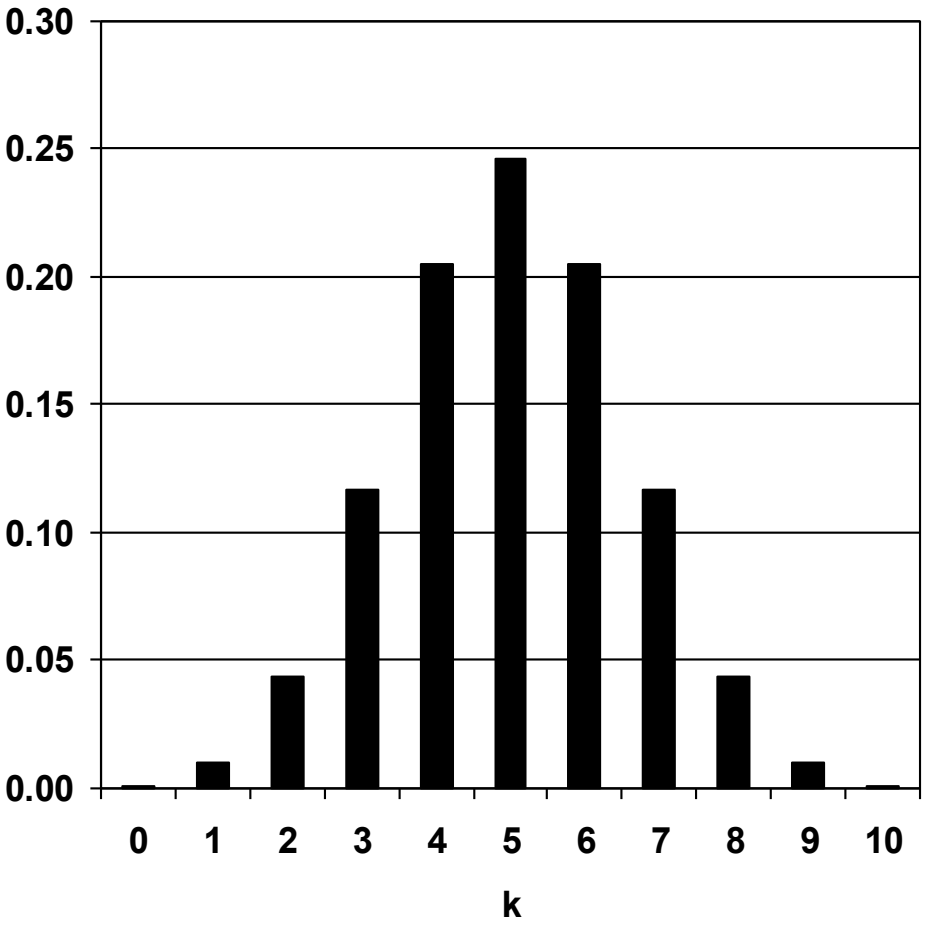
- Gọi p là tỉ lệ (xác suất) B > A
- Giả thuyết vô hiệu (H0): P = 0.5
- Giả thuyết chính (H1): B > A
- Dưới giả định H0 là đúng, chúng ta có thể tính xác suất có 0, 1, 2, v.v. bệnh nhân với kết qỉa B > A:

$$\Pr(k) = \binom{10}{k} p^k (1-p)^{10-k}$$

Tính toán ...

Xác suất có k = 0, 1, 2, 3, ..., 10 bệnh nhân với kết quả B > A

k	Pr(k)
0	0.0009765625
1	0.009765625
2	0.04394531
3	0.1171875
4	0.2050781
5	0.2460938
6	0.2050781
7	0.1171875
8	0.04394531
9	0.009765625
10	0.0009765625

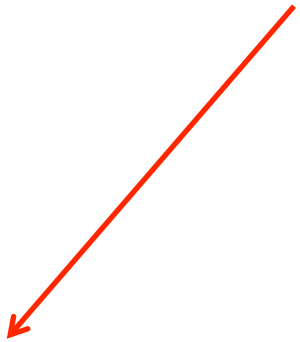


Tính toán ...

Xác suất có k = 0, 1, 2, 3, ..., 10 bệnh nhân với kết quả B > A

k	Pr(k)
0	0.0009765625
1	0.009765625
2	0.04394531
3	0.1171875
4	0.2050781
5	0.2460938
6	0.2050781
7	0.1171875
8	**0.04394531**
9	**0.009765625**
10	**0.0009765625**
P(k>=8)	**0.054687**

P(có ít nhất 8 bệnh nhân B>A) = 5.5%



P cỡ nào là "có ý nghĩa thống kê"

- *Statistically significant* – "có ý nghĩa thống kê"
- Fisher (1925) đề xuất $P < 0.05$ để tuyên bố ***statistically*** significant
- Ngưỡng này ngay sau đó được lấy làm chuẩn để đánh giá một giả thuyết, một khám phá (*Fisher không có ý định đó*!)

Những "ingredients" của một kiểm định thống kê

- Giả thuyết vô hiệu (null hypothesis) – không có khác biệt, không có tương quan, v.v.
- Dữ liệu (data) – test statistic (ie giá trị của t-test, Chi-square, hệ số tương quan, etc)

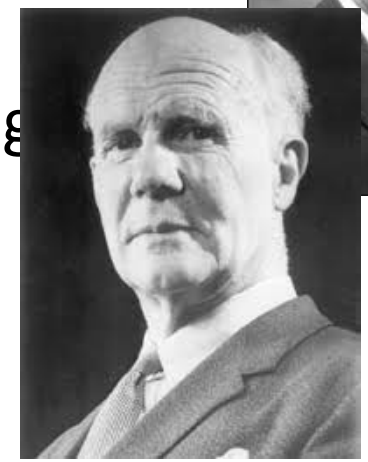
P-value = P(data | null hypothesis)

Test of hypothesis
(kiểm định giả thuyết)

Hai mô thức kiểm định

- **Test of significance** (Ronald A. Fisher)

  Thống kê là một khoa học *qui nạp (inductive inference)*: kết luận đi từ mẫu và áp dụng cho quần thể, từ nhỏ đến lớn.



- **Test of hypothesis** (Jerzy Neyman và Egon Pearson)

  NP bác bỏ khái niệm qui nạp trong thống kê; thống kê là một công cụ để "making decisions and guiding behavior"

Tranh luận giữ Fisher, Neyman - Pearson

- Neyman và Pearson phê phán phương pháp của Fisher (test of significance)
- Fisher "trả đũa"
- Một cuộc tranh luận kéo dài trên 10 năm và với nhiều ... bài báo khoa học trên *Biometrika*

Test of hypothesis (Neyman-Pearson)

- Phát biểu 2 giả thuyết, H_0 and H_1
- Quyết định α (xác suất bác bỏ H_1 nếu H_0 là đúng) và and β (xác suất bác bỏ H_0 nếu H_1 là đúng)
- Nếu dữ liệu (data) nằm trong vùng bác bỏ (rejection region) H_0, chấp nhận H_1; còn không thì chấp nhận H_0.

Nguyên lí: "in the long run of experience, we shall not be too often wrong"

Mô hình khoa học hiện hành: hỗn hợp giữa Fisher và NP

Propose a hypothesis – H1

Propose a null hypothesis – H0

Collect the data – D

Compute the probability of obtaining the finding – P(D | H0)

If P(D | H0) < 0.05, reject H0, accept H1

Logic đằng sau mô hình hiện hành

1. Nếu H_0 đúng, thì dữ liệu không thể xảy ra (mệnh đề 1)
2. Dữ liệu xảy ra trong thực tế (mệnh đề)
3. Do đó, H_0 không thể đúng (Kết luận 1)
4. Hoặc H_0 hoặc H_1 đúng
5. H_0 không đúng (mệnh đề 3)
6. Do đó, H_1 phải đúng (Kết luận 2)

Hiểu chẩn đoán để hiểu trị số P

Tiêu chuẩn vàng

Test	Bệnh	Không bệnh	
+ve	a Dương tính thật (true positive)	**b** **Dương tính giả** **(false positive)**	**a+b**
-ve	**c** **Âm tính giả** **(false negative)**	d Âm tính thật (True negative)	**c+d**
	a+c	**b+d**	

Hiểu chẩn đoán để hiểu trị số P

Tiêu chuẩn vàng

Test	Bệnh	Không bệnh	
+ve	a Dương tính thật (true positive)	**b** **Dương tính giả** **(false positive)**	**a+b**
-ve	**c** **Âm tính giả** **(false negative)**	d Âm tính thật (True negative)	**c+d**
	a+c	**b+d**	

- **Độ nhạy**: nếu có bệnh, xác suất có kết quả xét nghiệm dương tính là bao nhiêu?
- **Dương tính giả**: nếu KHÔNG có bệnh, xác suất có kết quả xét nghiệm dương tính là bao nhiêu?

Trong nghiên cứu khoa học

Hiệu quả (efficacy, liên quan)

Statistical Test	Có	Không
+ve	$1-\beta$ Power (Độ nhạy)	α type I error (false positive)
-ve	β (type II error)	$1-\alpha$

Chẩn đoán và nghiên cứu khoa học

Tiêu chuẩn vàng

Test	Bệnh	Không bệnh
+ve	a Dương tính thật (true positive)	b Dương tính giả (false positive)
-ve	c Âm tính giả (false negative)	d Âm tính thật (True negative)

Hiệu quả (efficacy, liên quan)

Statistical Test	Có	Không
+ve	1-b Power (Độ nhạy)	**a** **type I error** **(false positive)**
-ve	**b** **(type II error)**	1-a

Ý nghĩa của trị số P là gì?

- Xác suất mà kết quả kiểm định (t-test, Ki bình phương, hệ số tương quan, v.v.) **NẾU giả thuyết vô hiệu là đúng**
- **Giả thuyết vô hiệu** (nulll hypothesis): không có hiệu quả, không có liên quan, v.v.
- **Nếu thuốc không có hiệu quả**, xác suất quan sát kết quả hiện tại là bao nhiêu?
- Gần giống với *tỉ lệ dương tính giả* (nhưng không phải!)

Kiểm tra hiểu biết của bạn

"*Bisphosphonates giảm nguy cơ gãy xương 45% (*z = 2.15*; P = 0.03)*"

Phát biểu đó có nghĩa là:

- Xác suất thuốc không có hiệu quả (giả thuyết vô hiệu) là 3%
- Xác suất thuốc có hiệu quả (giả thuyết đảo) là 97%

Trị số P không phải là

- Xác suất mà **giả thuyết vô hiệu là đúng**
- Trị số P **không** phản ảnh mức độ khả dĩ của một giả thuyết

- Xác suất thuốc không có hiệu quả (giả thuyết vô hiệu) là 3%
- Xác suất thuốc có hiệu quả (giả thuyết đảo) là 97%

Kiểm tra hiểu biết của bạn

"*Bisphosphonates giảm nguy cơ gãy xương 45% (*z = 2.15; *P = 0.03)*"

Phát biểu đó có nghĩa là:

Nếu thuốc không có hiệu quả (giả thuyết vô hiệu là đúng), thì giá trị z ≥ 2.15 xảy ra là 3%.

Vấn đề của trị số P

Những vấn đề của trị số P

- Khó hiểu – vấn đề logic
- Không cho chúng ta biết về **tầm ảnh hưởng** (effect size)
- Không cung cấp thông tin chúng ta muốn biết: *khả năng giả thuyết đúng là bao nhiêu*?
- Phụ thuộc vào cỡ mẫu
- Có thể bị **nhiễu** khi kiểm định nhiều giả thuyết

Vấn đề logic

- Proof by contradiction (chứng minh nghịch đảo)
  - Tiền đề 1: nếu giả thuyết vô hiệu đúng thì kết quả này không thể xảy ra
  - Tiền đề 2: kết quả xảy ra
  - Kết luận: do đó, giả thuyết vô hiệu không đúng

Vấn đề logic

- Thử lí giải …
  - Tiền đề 1: nếu ông Tuấn là người Việt thì ông có thể không phải là đại biểu Quốc hội
  - Tiền đề 2: Ông Tuấn là đại biểu Quốc hội
  - Kết luận: do đó, ông Tuấn không phải là người Việt

Trị số P và tầm ảnh hưởng

- Trong nghiên cứu, chúng ta cần biết tầm ảnh hưởng (effect size, hệ số tương quan, v.v.)
- Trị số P không nói gì về tầm ảnh hưởng
- Kết quả với trị số P = 0.01 **không hẳn** có nghĩa là tầm ảnh hưởng lớn hơn với kết quả với trị số P = 0.04.

Trị số P và giả thuyết

- Chúng ta muốn biết với dữ liệu D thì khả năng giả thuyết đúng là bao nhiêu

$$P(H \mid D) = ?$$

- Nhưng trị số P cung cấp thông tin ngược!

$$P(D \mid H)$$

Cỡ mẫu lớn dễ có ý nghĩa thống kê

Có thể xem qua 4 nghiên cứu kiểm tra giả thuyết P = 0.5

Nghiên cứu	Cỡ mẫu	Tỉ lệ	Trị số P
1	20	15 (0.75)	0.041
2	200	114 (0.57)	0.041
3	2000	1046 (0.525)	0.041
4	2,000,000	1,001,445 (0.5007)	0.041

Tóm lại ...

Ba mô thức thống kê

- Estimation – ước tính
- Test of significance – Kiểm định thống kê
- Test of hypothesis – Kiểm định giả thuyết

Trị số P

- Trị số P là thước đo của sự ngạc nhiên
- Trị số P có nhiều vấn đề về logic và tuỳ thuộc vào cỡ mẫu
- Trị số P không trực tiếp phản ảnh tính khả dĩ của giả thuyết
- **Không nên quá lệ thuộc vào trị số P!**
- **Nên dùng ước số và khoảng tin cậy 95% (hoặc trường phái Bayes)**