

COURSE

Kiểm định giả thuyết thống kê

Lớp phân tích thống kê cơ bản

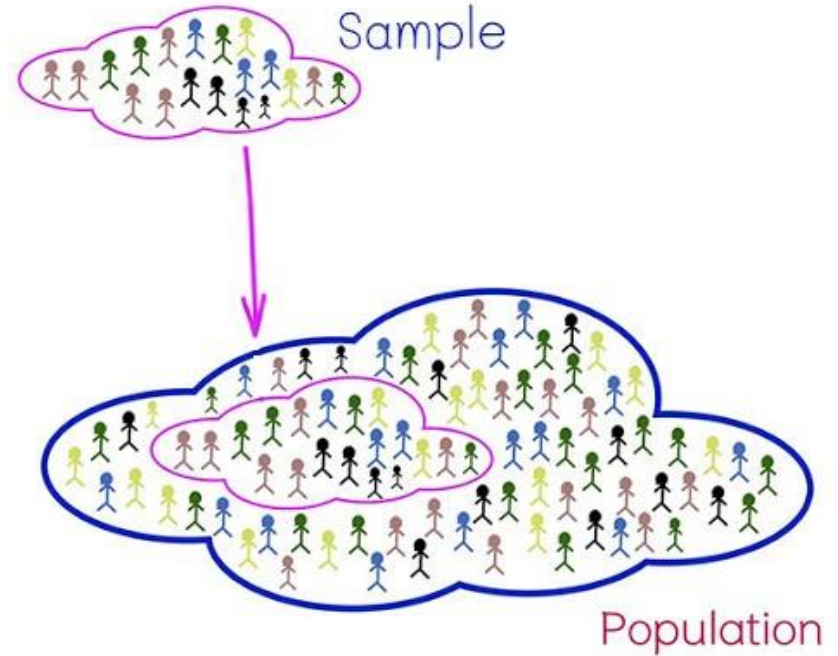
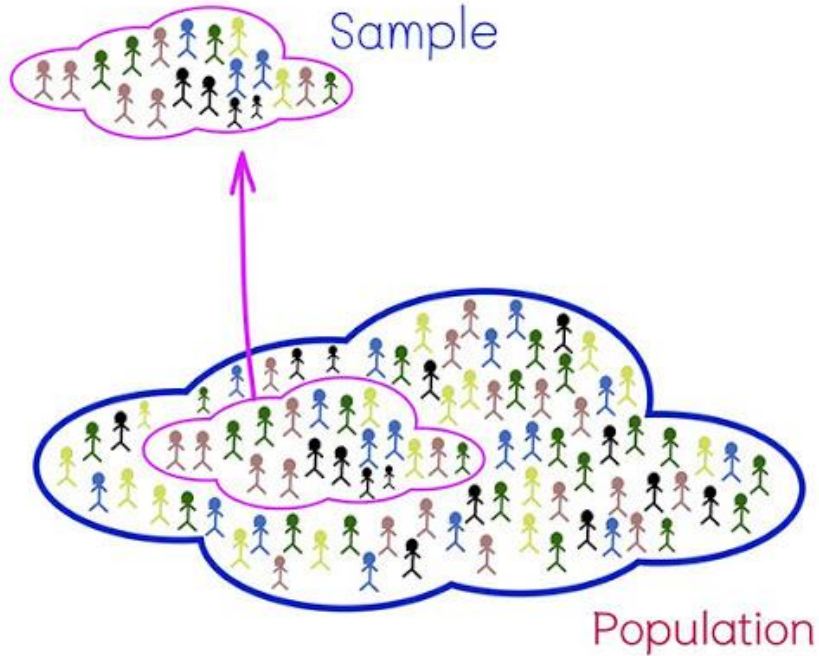
Khương Quỳnh Long
Hà Nội, 06-08/06/2020

Nội dung

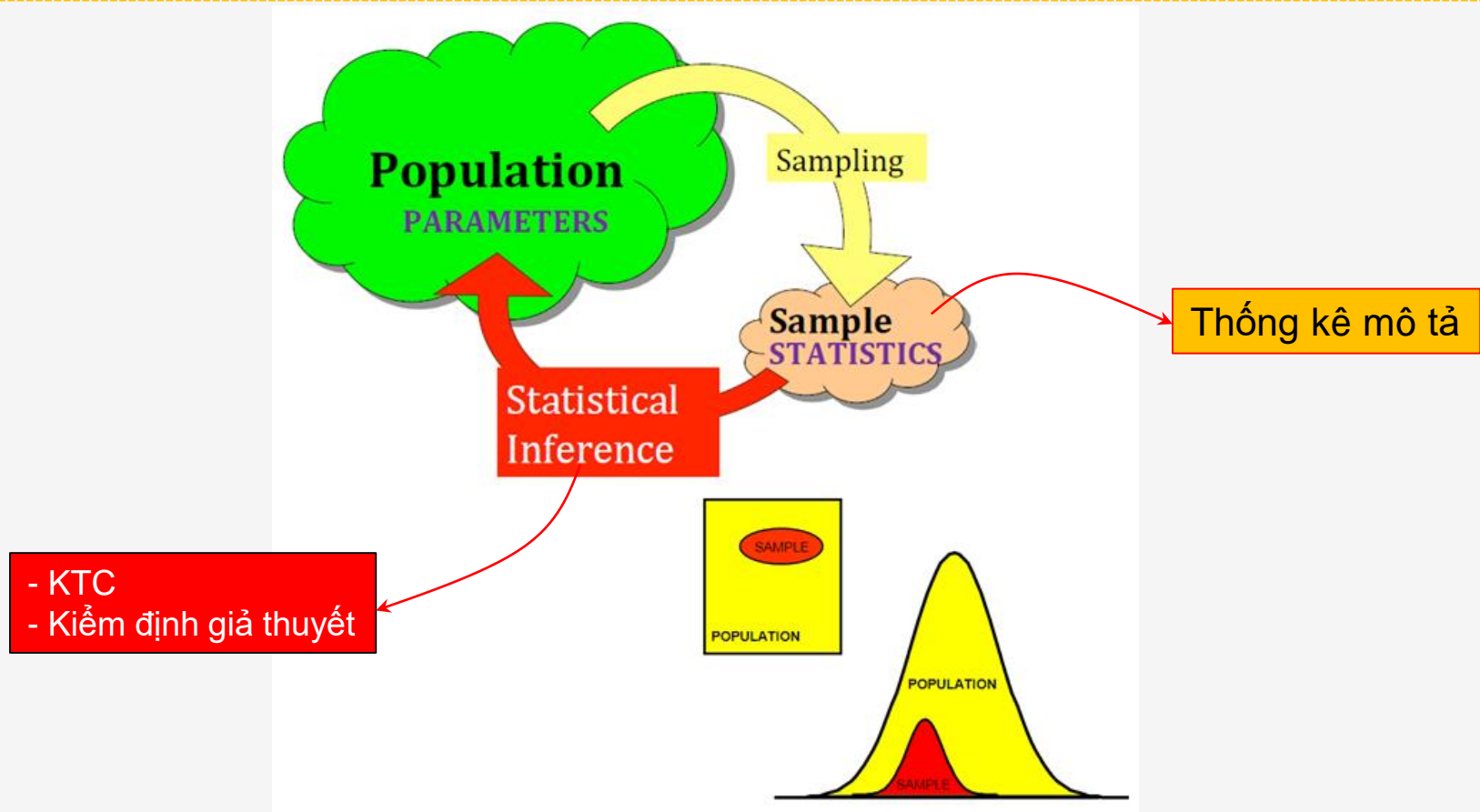
- Quần thể & mẫu
- Phân bố chuẩn & định lý giới hạn trung tâm
- Kiểm định giả thuyết thống kê
- P-value và KTC 95%
- Sai lầm loại I, loại II và lực thống kê
- Phân loại biến số

Quản thẻ và mẫu

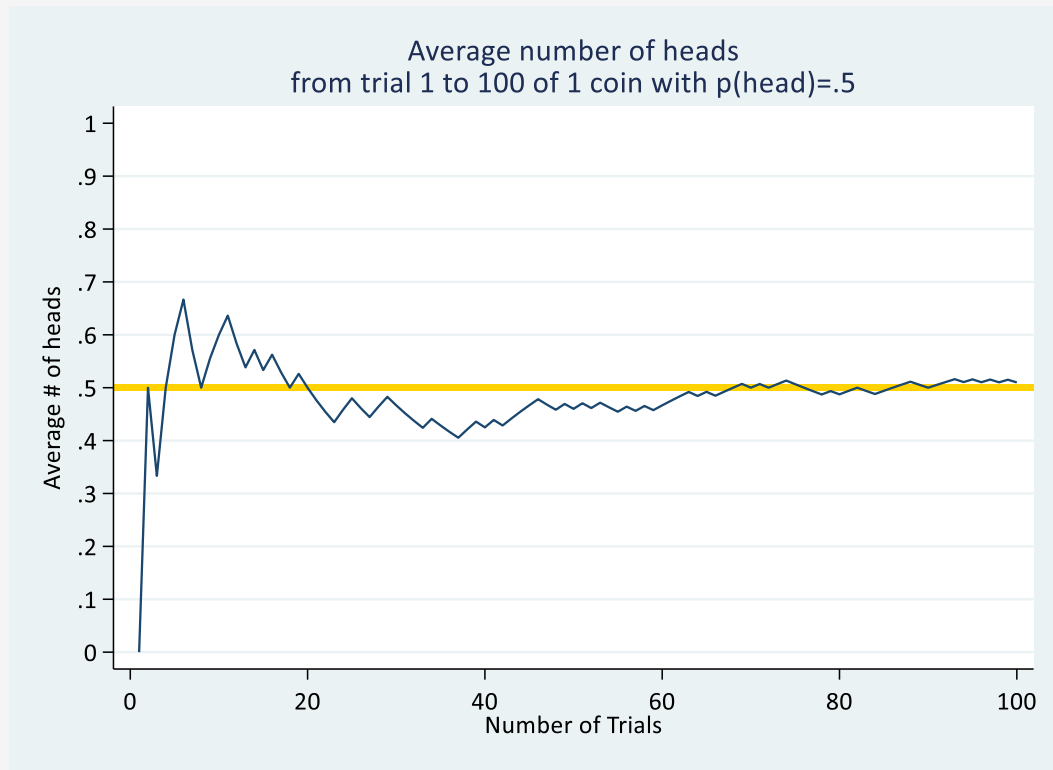
Quần thể & mẫu



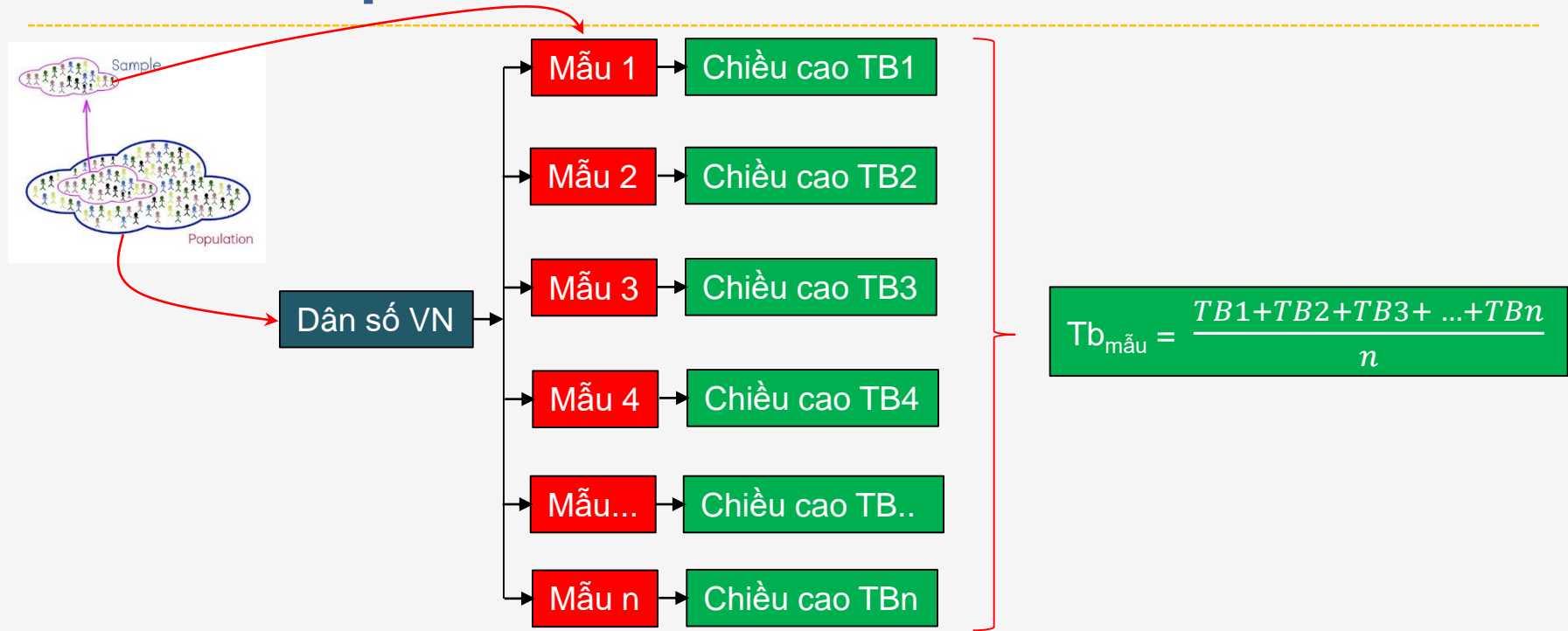
Quần thể & mẫu



Quy luật số lớn (Law of Large Numbers - LLN)



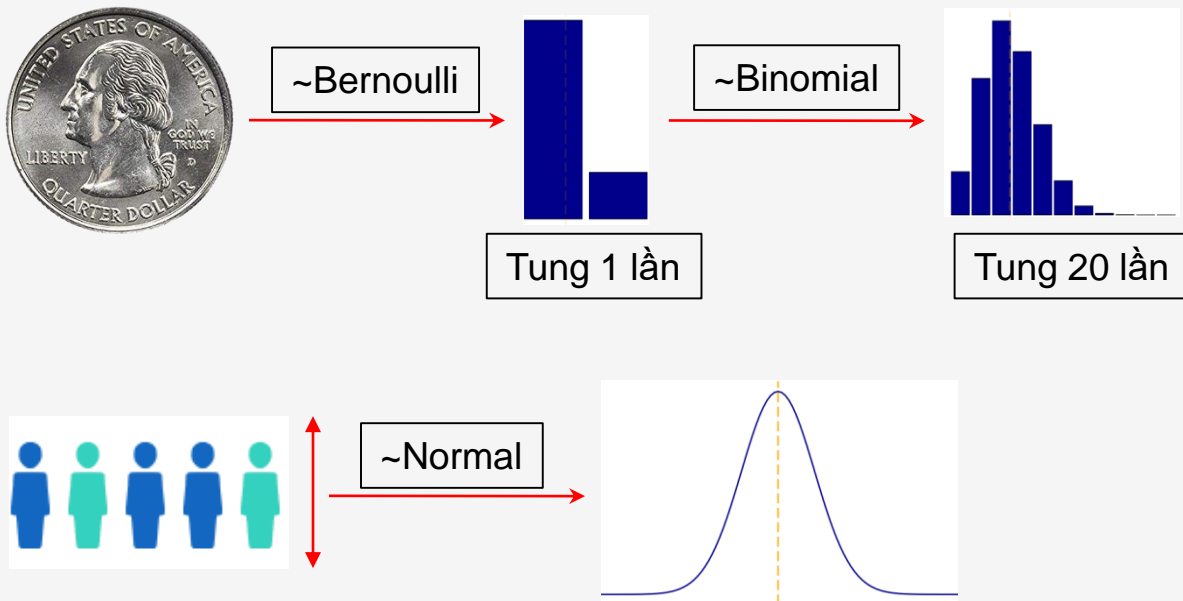
LLN - Ví dụ



Trung bình mẫu \rightarrow trung bình của dân số khi n càng lớn ($+\infty$)

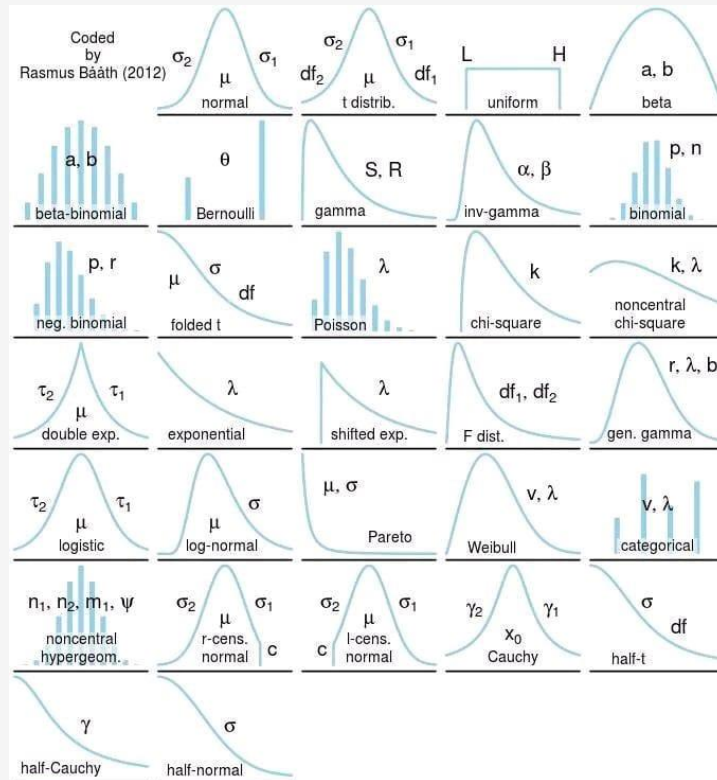
Phân bố thống kê

- Phân bố thống kê (Statistical distribution): là hàm toán học cung cấp xác suất của các kết cuộc có thể xảy ra



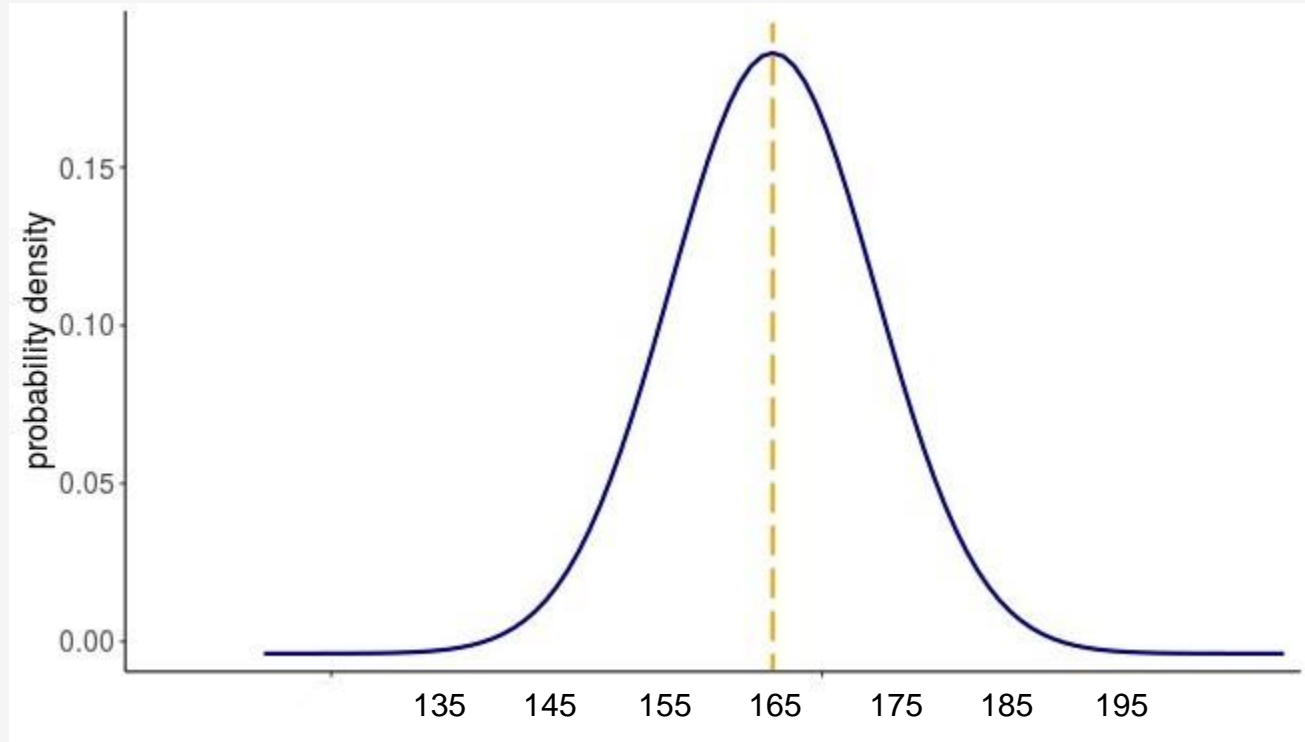
Phân bố thống kê

- Phân bố liên tục, rời rạc



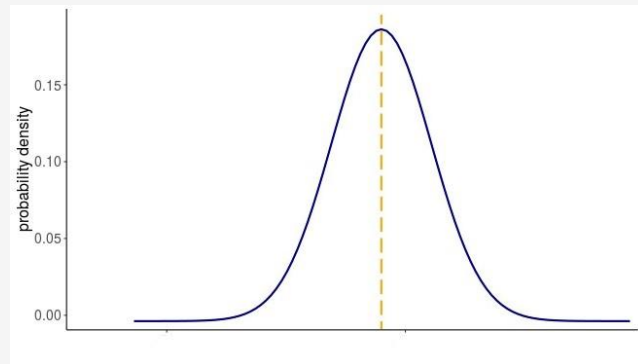
Phân bố chuẩn

Phân bố chiều cao

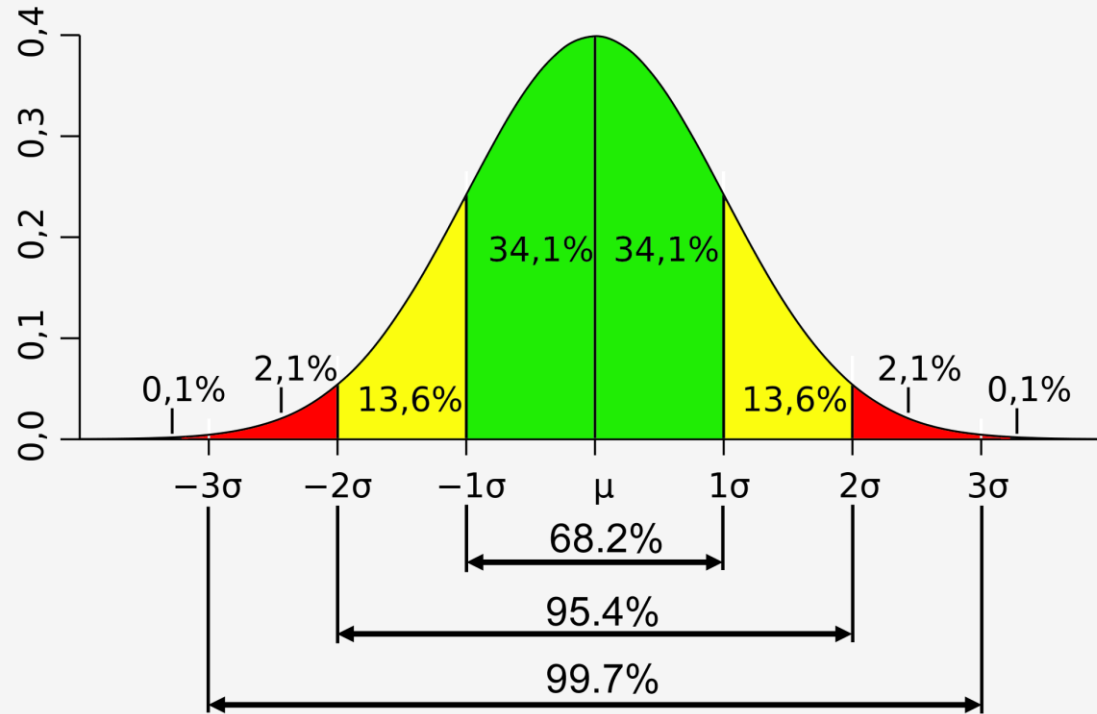


Phân bố chuẩn

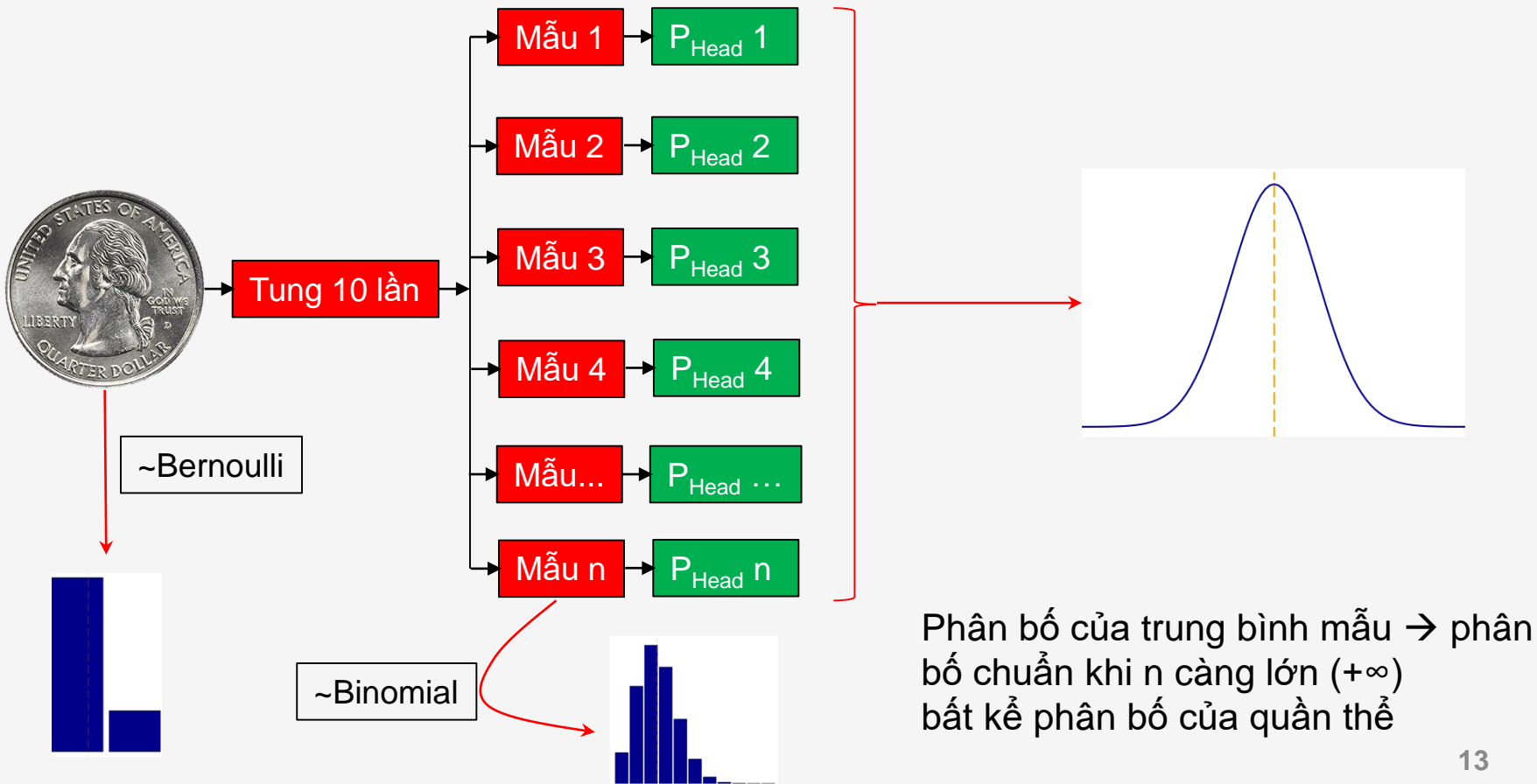
- Normal (Gaussian) distribution
- Hàm pdf $f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$
- $x \in (-\infty; +\infty)$
- 2 tham số:
 - ✓ Trung bình $\mu \in (-\infty; +\infty)$
 - ✓ Độ lệch chuẩn $\sigma \in (0; +\infty)$
- $\mu = 0, \sigma^2 = 1 \rightarrow$ Standardized normal distribution
- $Z \sim N(0, 1)$



Phân bố chuẩn



Định lý giới hạn trung tâm (CLT)



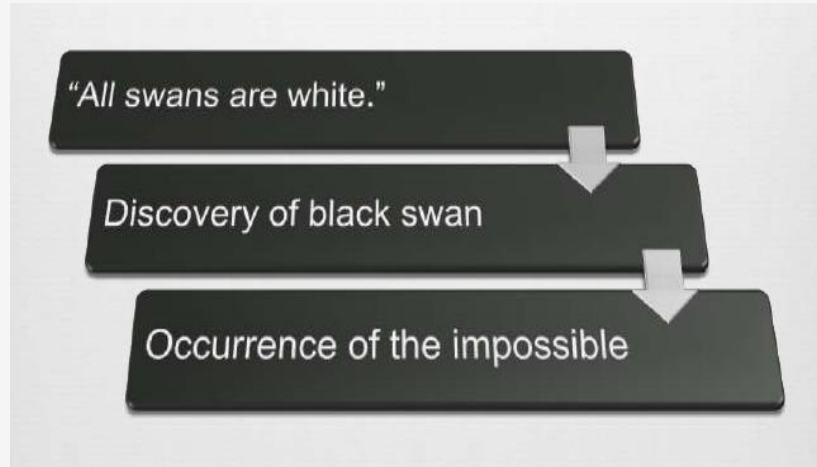
Định lý giới hạn trung tâm (CLT)

- Phân bố của trung bình mẫu \rightarrow phân bố chuẩn khi n càng lớn ($+\infty$) bất kể phân bố của quần thể
- Vai trò quan trọng
 - ✓ Ước lượng khoảng tin cậy
 - ✓ Kiểm định giả thuyết thống kê

Kiểm định giả thuyết

Quá trình kiểm định ý nghĩa thống kê

- Được giới thiệu bởi Ronald Fisher vào 1920s
- Dựa trên triết lý phản nghiệm (falsificationism)
- Không bao giờ chứng minh được một giả thuyết
- Chỉ có thể bác bỏ giả thuyết



Quá trình kiểm định ý nghĩa thống kê

1. Xây dựng giả thuyết vô hiệu (Null) H_0
✓ → giả thuyết thay thế H_A
2. Chọn lựa kiểm định thích hợp
3. Tính giá trị thống kê của số liệu thu thập được
4. Tính xác suất quan sát được biến cố và những trường hợp “hiếm hơn” (observed and more extreme results) khi H_0 đúng $P(D|H_0)$ hay **P value**
5. P value càng nhỏ → càng cho thấy bằng chứng bác bỏ H_0
→ Nếu P value đủ nhỏ chúng ta kết luận $P(H_0)$ nhỏ và bác bỏ giả thuyết H_0 . Ngưỡng ý nghĩa thường được lấy = 0.05

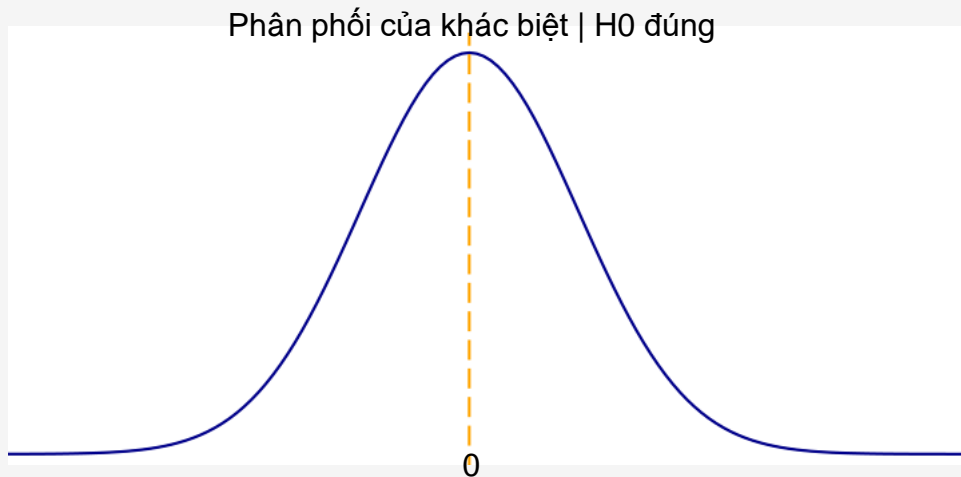
Ví dụ



- Tung đồng xu (lạ) 10 lần, xuất hiện 8 lần head
 - Đồng xu cân bằng?
- Kiểm định giả thuyết

Ví dụ

- B1: giả thuyết H_0
 - ✓ Đồng xu cân bằng (không có khác biệt giữa head và tail), hay $P_{\text{head}} = 0.5$
 - ✓ Giả thuyết thay thế $P_{\text{head}} \neq 0.5$
 - ✓ Phân bố của sự khác biệt (\sim CLT)



Ví dụ

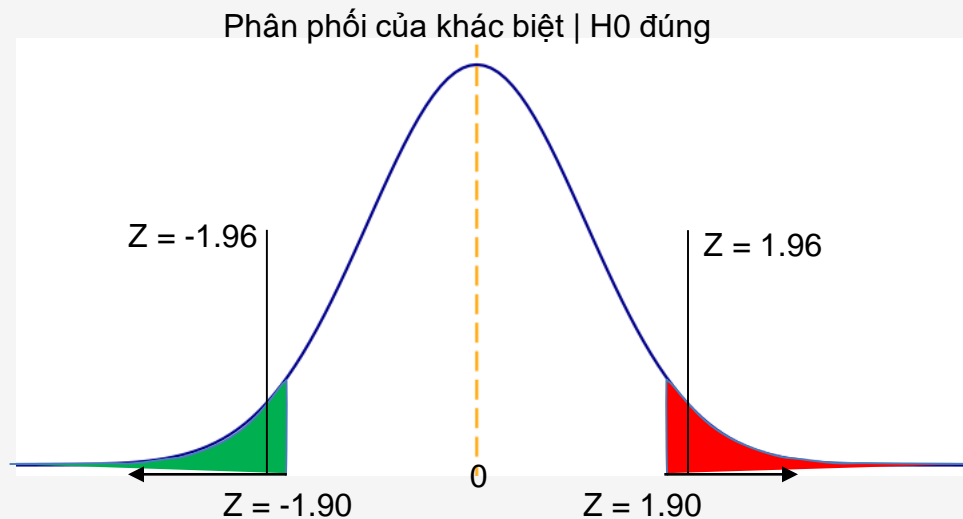
- B2: lựa chọn phép kiểm phù hợp
 - ✓ Kiểm định sự khác biệt cho 1 tỷ lệ
- B3: tính toán giá trị thống kê (Z)
 - ✓ $Z = \text{diff}/\text{standard error}$
$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$
 - ✓ $Z = (0.8-0.5)/\text{sqrt}(0.5*(1-0.5)/10) = 1.90$

Ví dụ

B4: Tính P-value: $P(Z > z \mid H_0)$

✓ P 2-tails = $P(Z > |1.90|) = \text{xanh} + \text{đỏ} = 0.0578$

✓ P one-tail = $P(Z > 1.90) = \text{đỏ} = 0.0289$



Ví dụ

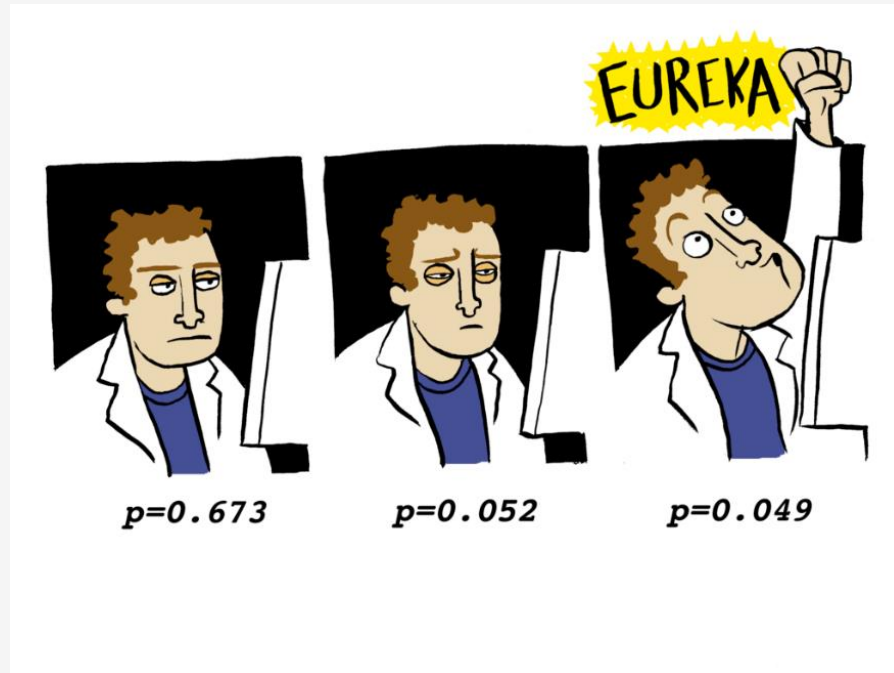
- B5: Kết luận
- Dựa vào $P\text{-value} = 0.0578$
 - không đủ bằng chứng bác bỏ H_0 (không có ý nghĩa thống kê)
 - chấp nhận H_0
 - đồng xu cân bằng!

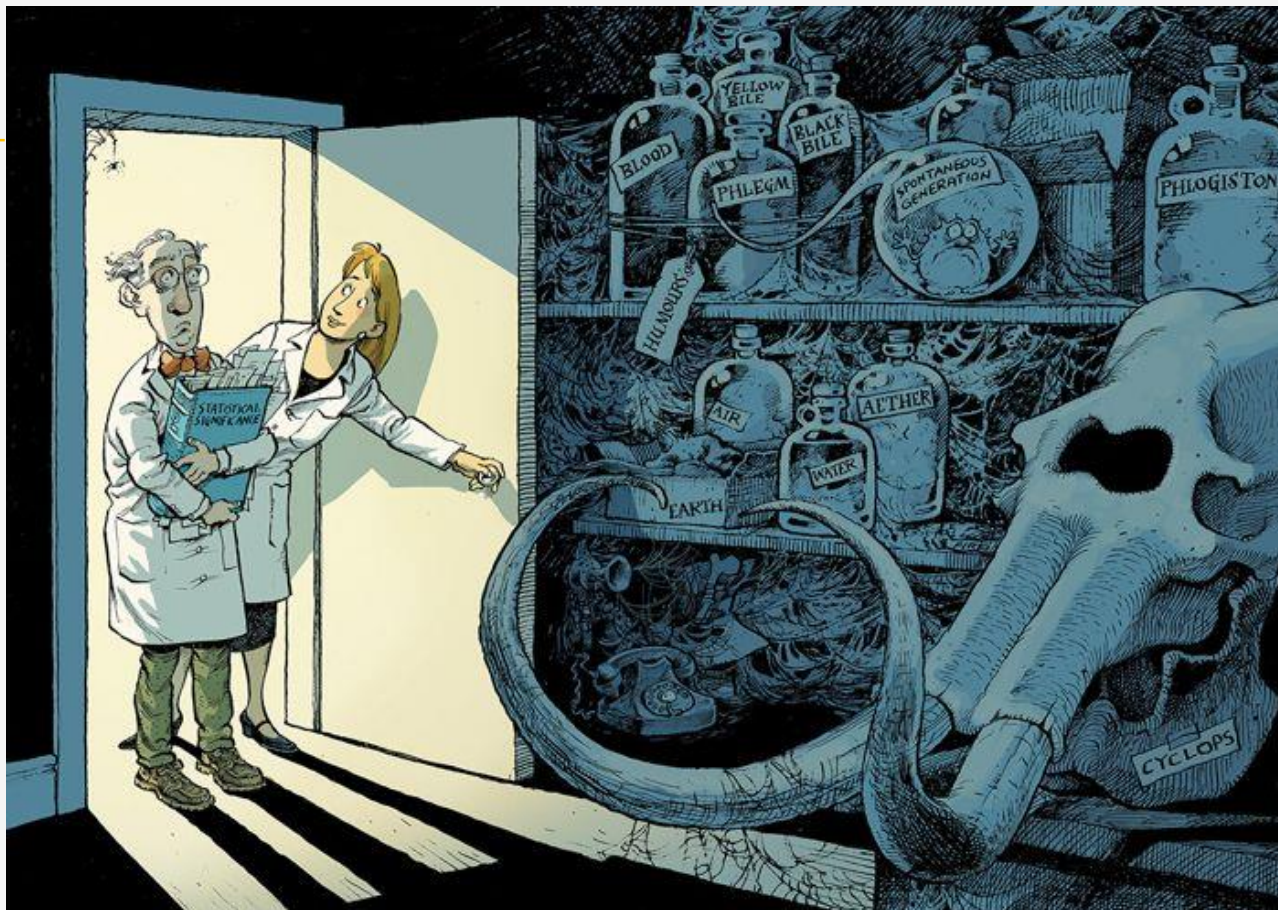
Ví dụ

- Lặp lại ví dụ, nhưng tun 100 lần, xuất hiện head 80 lần ($P = 0.8$)
- Kết luận?
- Nhận xét?

Thận trọng với “Ý nghĩa thống kê”

Fishing!





Abandon Statistical Significance: <https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1527253>

Retire Statistical Significance: <https://www.nature.com/articles/d41586-019-00857-9>

Moving to a World Beyond “ $p < 0.05$ ”: <https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913>

Thận trọng với “Ý nghĩa thống kê”

- P-value phụ thuộc cỡ mẫu
 - Ý nghĩa thống kê \neq ý nghĩa thực tế
- Tham khảo thêm các “P-value fallacy”

“The difference between ‘significant’ and ‘not significant’ is not itself statistically significant.”¹

¹Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett (2019) Abandon Statistical Significance, The American Statistician, 73:sup1, 235-245

Khoảng tin cậy

Khoảng tin cậy 95%

Nghiên cứu trên 10,000 người,
tỉ lệ ung thư là 10% (KTC 95% = 8% – 12%)

Nghĩa là?

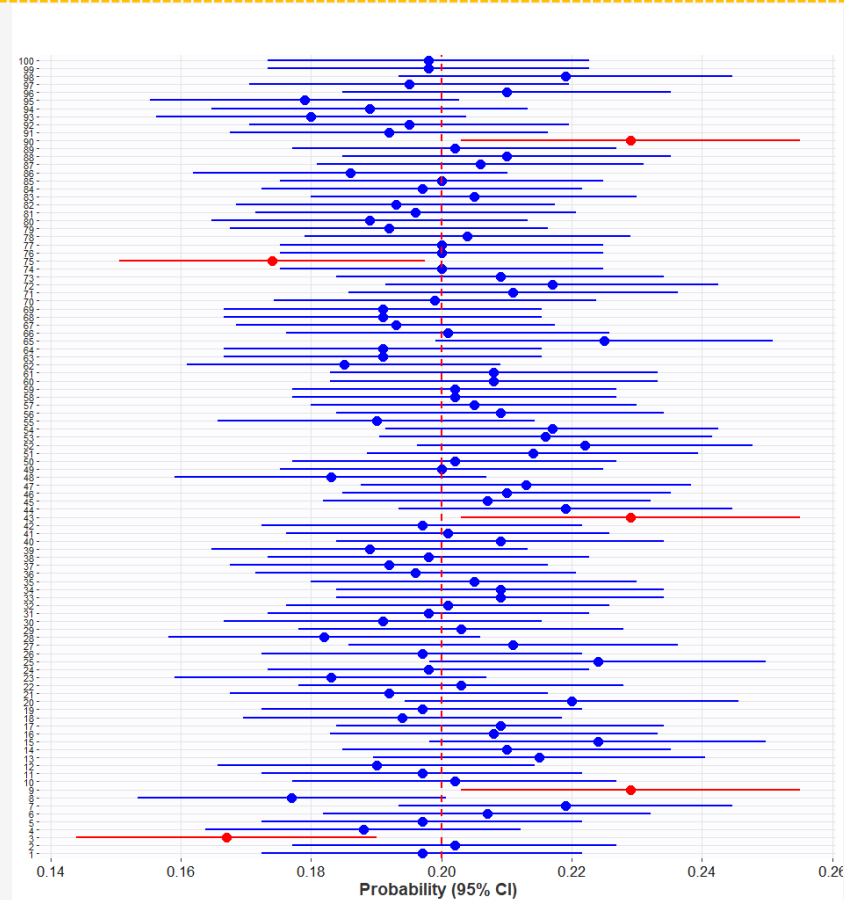
Khoảng tin cậy 95%

- 95% khoảng tin cậy sẽ chứa giá trị thật của dân số?
- Xác suất để giá trị thật của dân số nằm trong khoảng này là 95% ?

Khoảng tin cậy 95%

- Theo Frequentist, giá trị thật của dân số là **thật**, nhưng **không biết** (unknown) và **cố định** (fixed).
 - Nếu lặp lại nghiên cứu tương tự n lần, **95% số n lần** sẽ tạo thành khoảng chứa giá trị thật của dân số
 - 95%CI là một **quá trình** “long-run” (95% của n lần), **không cụ thể** cho một mẫu (nghiên cứu)
- ➔ Cho một nghiên cứu cụ thể, giá trị thật của dân số **có thể nằm trong 95%CI hoặc không !**

Khoảng tin cậy 95%



Sai lầm loại I, II & lực thống kê

Xác suất sai lầm loại 1, 2 và lực thống kê (power)

- Thống kê là khoa học dựa vào nguyên lý xác suất
- Kết luận đều được diễn giải trong một sai số quy ước (chấp nhận được)

Kết luận từ phương pháp thống kê		Thực tế	
		Thuốc A = placebo (H_0)	Thuốc A > placebo (H_A)
	Bác bỏ H_0	Trường hợp 1	Trường hợp 2
	Chấp nhận H_0	Trường hợp 3	Trường hợp 4

Xác suất sai lầm loại 1 (α)

- Trường hợp 1
- Xác suất bác bỏ H_0 khi H_0 đúng
- Xác suất kết luận thuốc A > placebo trong khi thực tế thuốc A = placebo
- Khả năng kết luận 2 yếu tố có liên quan, có khác biệt (bác bỏ H_0) trong khi nó thực sự không có liên quan, không có khác biệt
- “Dương tính giả”
- Thường được chọn = 5%

Kết luận từ phương pháp thống kê		Thực tế	
		Thuốc A = placebo (H_0)	Thuốc A > placebo (H_A)
	Bác bỏ H_0	Trường hợp 1	Trường hợp 2
	Chấp nhận H_0	Trường hợp 3	Trường hợp 4

Xác suất sai lầm loại 2 (β)

- Trường hợp 4
- Xác suất chấp nhận H_0 khi H_0 sai
- Xác suất kết luận thuốc A = placebo trong khi thực tế thuốc A > placebo
- Khả năng kết luận 2 yếu tố không liên quan, không khác biệt (chấp nhận H_0) trong khi nó thực sự có liên quan, có khác biệt
- “Âm tính giả”
- Thường được chọn = 20%

Kết luận từ phương pháp thống kê		Thực tế	
		Thuốc A = placebo (H_0)	Thuốc A > placebo (H_A)
	Bác bỏ H_0	Trường hợp 1	Trường hợp 2
	Chấp nhận H_0	Trường hợp 3	Trường hợp 4

Lực thống kê (statistical power)

- Trường hợp 2
- $1 - \beta$ – xác suất sai lầm loại 2 (β)
- Xác suất bác bỏ giả thuyết H_0 (chấp nhận H_A) khi giả thuyết H_0 sai
- Xác suất kết luận thuốc A > placebo khi thực tế thuốc A > placebo
- “Dương tính thật”

Kết luận từ phương pháp thống kê		Thực tế	
		Thuốc A = placebo (H_0)	Thuốc A > placebo (H_A)
	Bác bỏ H_0	Trường hợp 1	Trường hợp 2
	Chấp nhận H_0	Trường hợp 3	Trường hợp 4

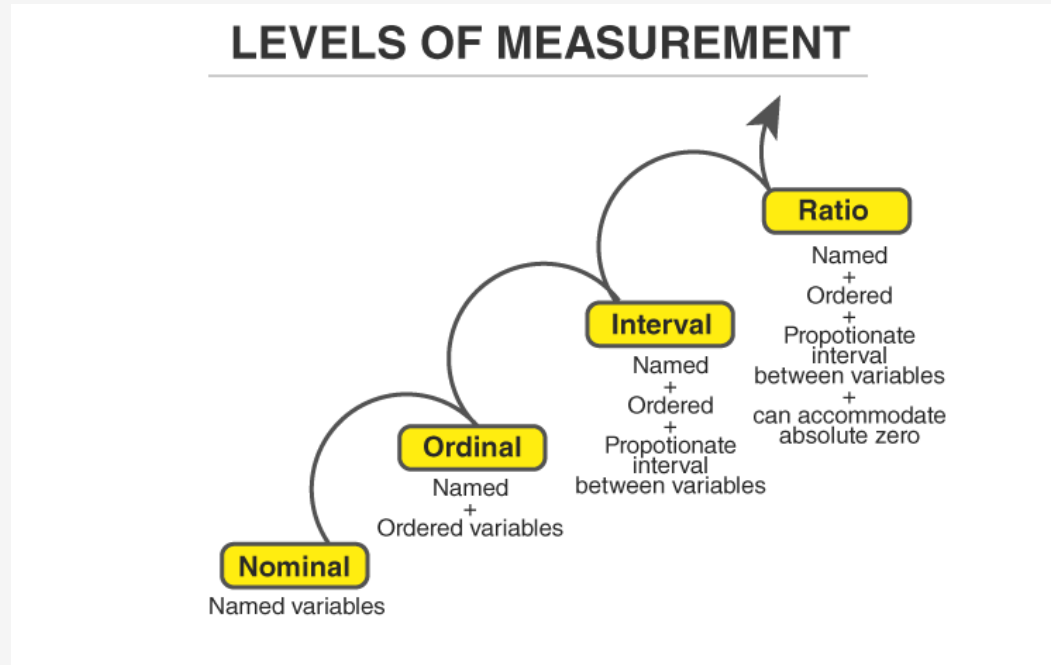
Tóm tắt

	Không có bệnh	Có bệnh
Chẩn đoán có bệnh $P < \alpha$	Chẩn đoán sai Dương tính giả Sai lầm loại 1	Chẩn đoán đúng Dương tính thật Lực thống kê
Chẩn đoán không có bệnh $P > \alpha$	Chẩn đoán đúng Âm tính thật	Chẩn đoán sai Âm tính giả Sai lầm loại 2

Phân loại biến số

Phân loại biến số

- Có bao nhiêu loại biến số?



Nội dung đã học

- Quần thể & mẫu
- Phân bố chuẩn & định lý giới hạn trung tâm
- Kiểm định giả thuyết thống kê
- P-value và KTC 95%
- Sai lầm loại I, loại II và lực thống kê
- Phân loại biến số