COURSE

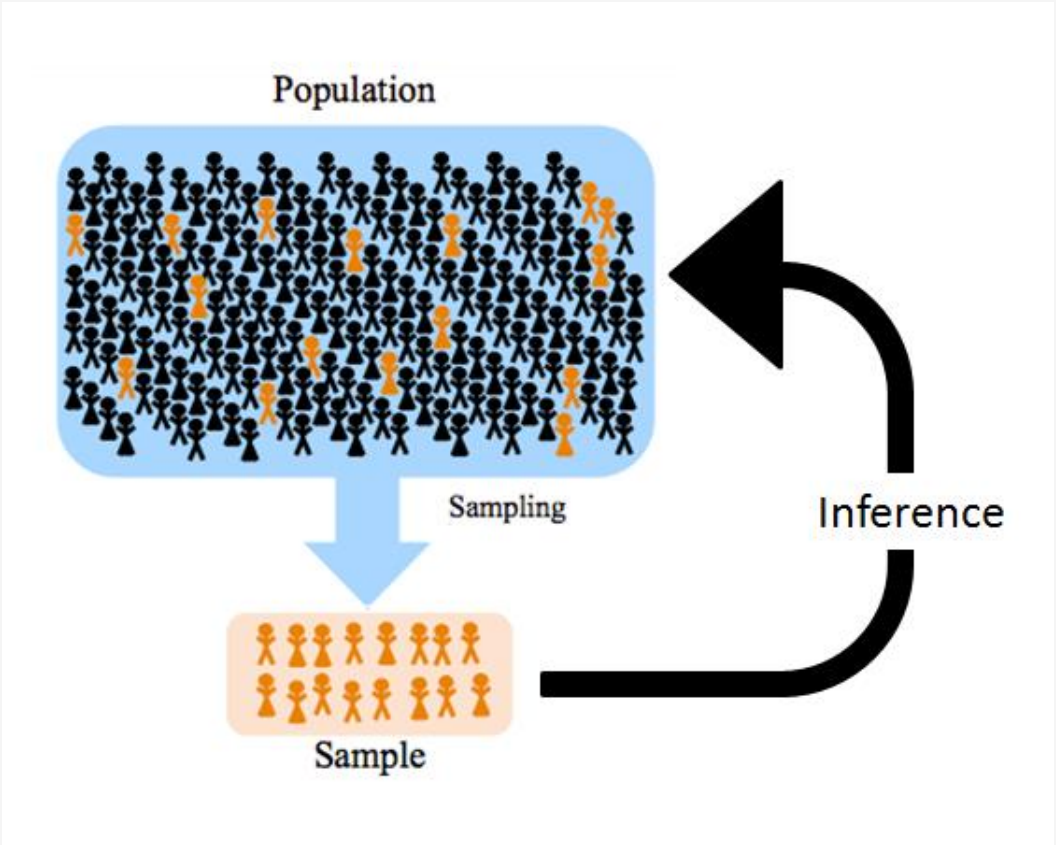
# Thống kê mô tả

## Lớp phân tích thống kê cơ bản

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

# Nội dung

- Các chỉ số thống kê mô tả
- Giới thiệu phần mềm Stata
- Thống kê mô tả trên Stata


Population
Sampling
Inference
Sample

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Trung bình:** $\bar{x} = \frac{\sum x_i}{N}$
    - Ví dụ HATT: $\bar{x} = \frac{\Sigma x_i}{N} = \frac{120 + 125 + 130 + 135 + 150}{5} = 132$
  - **Độ lệch chuẩn:** mức độ phân tán của số liệu quanh giá trị trung bình

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N - 1}}$$

- Ví dụ:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N-1}}$$

$$= \sqrt{\frac{(120-132)^2 + (125-132)^2 + (130-132)^2 + (135-132)^2 + (150-132)^2}{5-1}}$$

$$= \sqrt{\frac{144+49+4+9+324}{4}} = \sqrt{\frac{530}{4}} = \quad 132{,}5 = 11{,}5$$

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Phương sai:** độ biến động của số liệu, bằng bình phương của độ lệch chuẩn

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - x)^2}{N-1}$$

  - **Sai số chuẩn:** độ sai lệch của ước lượng so với giá trị của dân số

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Ví dụ: 120, 125, 130, 135, 150 có độ lệch chuẩn là 11.5 thì sai số chuẩn là:

$$SE = \frac{11.5}{\sqrt{5}} = 5.15$$

  - **Khoảng tin cậy 95%:** nếu lặp lại nghiên cứu n lần, 95% số lần sẽ tạo thành khoảng chứa giá trị của dân số

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Khoảng tin cậy 95%:** nếu lặp lại nghiên cứu n lần, 95% số lần sẽ tạo thành khoảng chứa giá trị của dân số
    - **Phương pháp Z:**

$$KTC = \bar{X} \pm Z_{1-\frac{\alpha}{2}} \text{ x SE}$$

    - Ví dụ: 120, 125, 130, 135, 150 có SE = 5.15, trung bình = 132, tra bảng Z với alpha = 0.05 → Z = 1.96 thì có KTC 95% là

  – Giới hạn **trên**:132 + 1.96 x 5.15 = 142.09

  – Giới hạn **dưới**: 132 - 1.96 x 5.15 = 121.91

  → KTC 95% = 121.91 - 142.09

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Trung vị:** giá trị nằm giữa chia số liệu thành 2 phần bằng nhau
    - Ví dụ 1: 120, 135, 125, 150, 130
    - Sắp xếp lại thứ tự: 120, 125, 130, 135, 150 → trung vị =130
    - Ví dụ 2: 120, 135, 125, 150, 130, 128
    - Sắp xếp lại thứ tự: 120, 125, 128, 130, 135, 150 → trung vị = (128 + 130) / 2 = 129

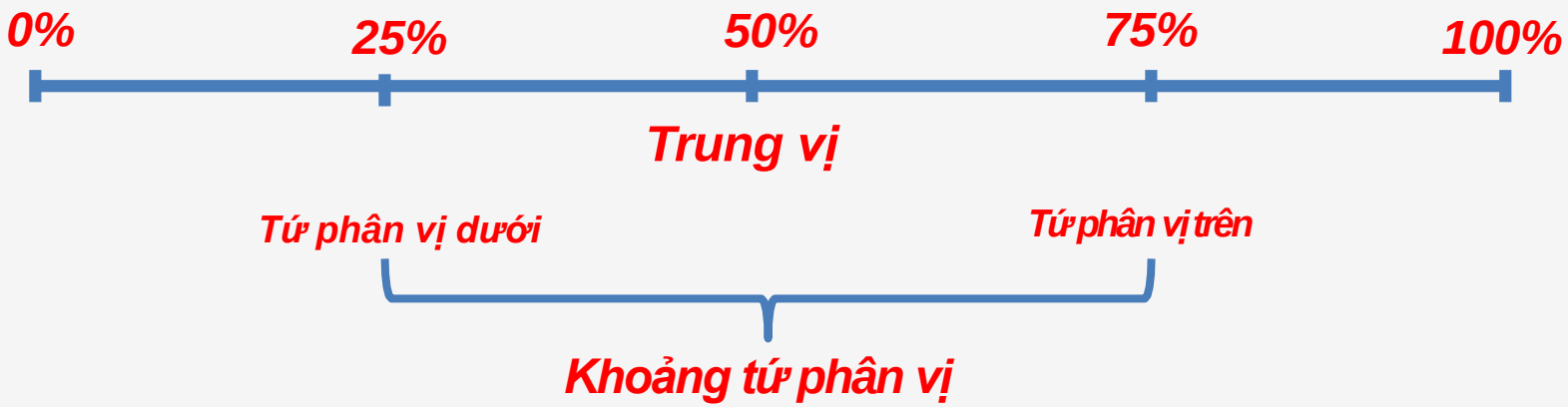
# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Khoảng tứ vị:** là khoảng cách của trung vị phần trên và trung vị phần dưới
    - Trung vị của phần trên của số liệu được gọi là tứ phân vị trên (upper quartile)
    - Trung vị của phần dưới số liệu được gọi là tứ phân vị dưới (lower quartile)
    - Ví dụ: 120, 125, 130, 135, 150.

    ✓ Chia số liệu làm 2 phần:
      - Phần 1: 120, 125, 130 → trung vị dưới: 125
      - Phần 2: 130, 135, 150 → trung vị trên: 135

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Trung vị và khoảng tứ phân vị**



# Các chỉ số thống kê mô tả

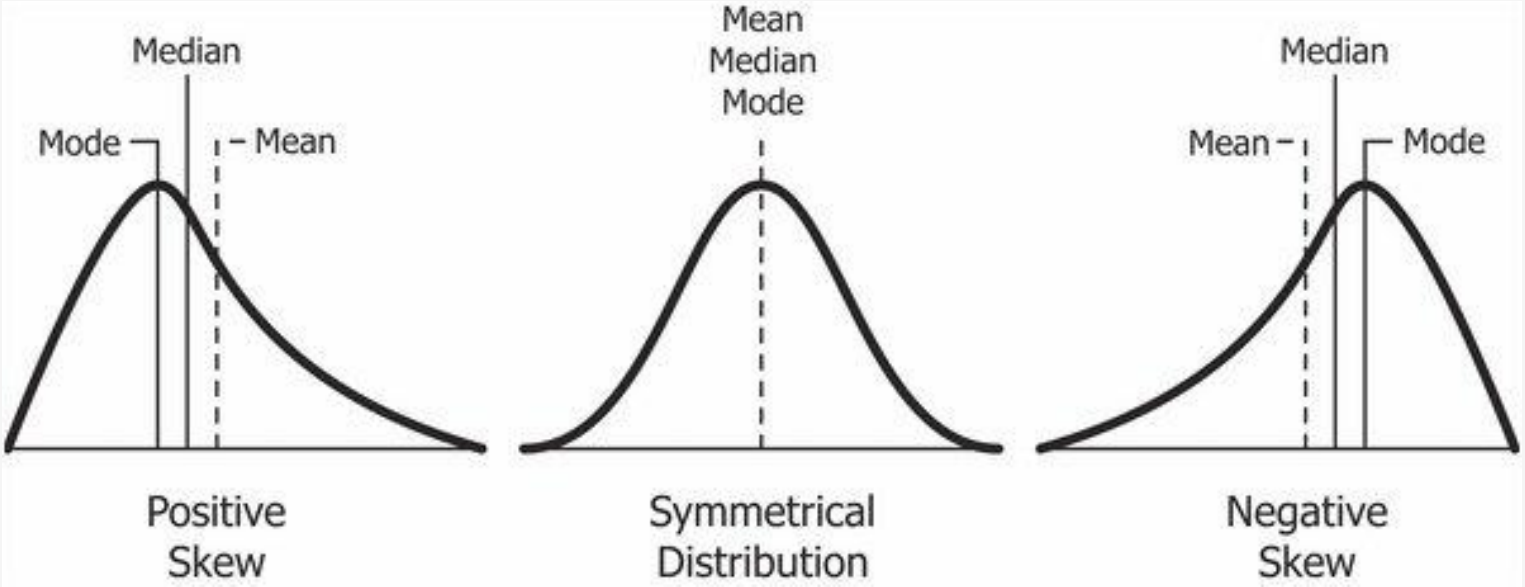
- Biến số định lượng
  - **Mode (yếu vị):** là giá trị xuất hiện phổ biến nhất (có tần suất cao nhất)
    - Ví dụ 1: 120, 125, 130, 135, 150 → không có mode
    - Ví dụ 2: 5, 5, 6, 7, 9 → mode là 5
    - Có thể không có mode, có thể có một mode hoặc hai hay nhiều mode

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Phạm vi:** khoảng giá trị từ giá trị nhỏ nhất (minimum) đến giá trị lớn nhất (maximum)
    - Ví dụ: 120, 125, 130, 135, 150 → phạm vi số liệu là 120 -150

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - **Phân phối bình thường**




**PP lệch phải** **PP bình thường** **PP lệch trái**

# Các chỉ số thống kê mô tả

- Biến số định lượng
  - So sánh trung bình và trung vị
    - Ví dụ: 4, 5, 3, 6, 7, 3, 5

    → Trung bình = 4.7; trung vị = 5

    - Ví dụ: 4, 5, 3, 6, 7, 43, 5

    → Trung bình = 10,42; trung vị = 5
  - Số liệu có phân phối bình thường sử dụng trung bình (độ lệch chuẩn, phạm vi)
  - Số liệu không có phân phối bình thường sử dụng trung vị (khoảng tứ phân vị, phạm vi)

# Các chỉ số thống kê mô tả

- Biến số định tính
  - **Tần số:** số quan sát
  - **Tỉ lệ:** số quan sát của đặc tính quan tâm chia cho tổng số quan sát
  - **Tỉ lệ cộng dồn**: số cộng dồn đến một mức giá trị của biến số (chỉ dùng cho biến thứ tự)

| Tình trạng sức khỏe | n | % | % cộng dồn |
|---|---|---|---|
| Rất tốt | 9 | 2.27 | 2.27 |
| Tốt | 73 | 18.39 | 20.65 |
| Bình thường | 258 | 64.99 | 85.64 |
| Yếu | 54 | 13.60 | 99.24 |
| Rất yếu | 3 | 0.76 | 100.00 |
| Tổng | 397 | 100 | |

# Các chỉ số thống kê mô tả

- Biến số định tính
  - Khoảng tin cậy 95%

$$p \pm 1.96\sqrt{p(1-p)/n}$$

  - Ví dụ: Quan sát 100 người thấy có 10 người ung thư

→Tỉ lệ ung thư: 10 :100 = 0.10 ~ 10%

- Giới hạn dưới $0.10-1.96\sqrt{0.10(1-0.10)/100} = 0.0412 \sim 4.12\%$
- Giới hạn trên $0.10+1.96\sqrt{0.10(1-0.10)/100} = 0.1588 \sim 15.88\%$

→ KTC 95%: 4.12% - 15.88%

# Các chỉ số thống kê mô tả

- Biến định tính
  - ✓ Tần số
  - ✓ Phần trăm
  - ✓ Phần trăm cộng dồn
  - ✓ KTC 95%

- Biến định lượng
  - ❖ Có phân phối bình thường
    - ✓ Trung bình
    - ✓ Độ lệch chuẩn
    - ✓ Phạm vi
  - ❖ Không có phân phối bình thường
    - ✓ Trung vị
    - ✓ Khoảng tứ phân vị
    - ✓ Phạm vi
  - ❖ KTC 95%

# Thống kê mô tả trên Stata

- Mở dữ liệu **sl7_huyetap.dta**
- Nhập vào lệnh **des**
- Thông tin thể hiện cần quan tâm bao gồm:
  - ***contain data from***: nơi lưu trữ file số liệu
  - ***obs***: Số đối tượng trong nghiên cứu
  - ***vars***: Số biến số trong nghiên cứu
  - ***Variable name***: tên biến số trong số liệu
  - ***Variable label***: nhãn của biến số
  - ***Sorted by***: sắp xếp số liệu theo biến số nào

```
. des

Contains data from D:\Dropbox\Long\Projects\Courses\Basic\Thống kê mô tả\sl7_huyetap.dta
  obs:           397
 vars:            19                          24 Apr 2012 16:11
```

| variable name | storage type | display format | value label | variable label |
|---|---|---|---|---|
| ma | float | %9.0g | | ma ca nhan |
| tuoi | float | %9.0g | | tuoi |
| gioitinh | float | %9.0g | sb3 | gioi tinh |
| cannang | float | %9.0g | | can nang |
| caotb | float | %9.0g | | chieu cao tb |
| suckhoe | float | %9.0g | sb15 | nhan xet ve suc khoe |
| ttthainghen | float | %9.0g | sb4 | tinh trang thai nghen |
| caoha | float | %12.0g | c | cao huyet ap |
| chieucao1 | float | %9.0g | | chieu cao lan 1 |
| chieucao2 | float | %9.0g | | chieu cao lan 2 |
| hatoida1 | float | %9.0g | | huyet ap toi da 1 |
| hatoida2 | float | %9.0g | | huyet ap toi da 2 |
| hatoithieu1 | float | %9.0g | | huyet ap toi thieu 1 |
| hatoithieu2 | float | %9.0g | | huyet ap toi thieu 2 |
| nhomtuoi | float | %9.0g | b | nhom tuoi |
| hatdtb | float | %9.0g | | huyet ap toi da tb |
| hatttb | float | %9.0g | | huyet ap toi thieu tb |
| bmi | float | %9.0g | | chi so khoi co the |
| beogay | float | %10.0g | g | tinh trang beo gay |

```
Sorted by:
```

# Thống kê mô tả trên Stata

- Biến định tính
  - Tần số
  - Tỉ lệ phần trăm
  - Tỉ lệ phần trăm cộng đồn (với biến thứ tự)
  - Khoảng tin cậy 95%
  - Stata
    - ***tab1 {các biến số}***
      - Ví dụ: `tab1 gioitinh suckhoe`
    - ***ci prop {biến nhị giá}*** *# cần mã hóa 0/1*
      - Ví dụ: `ci prop caoha`

# Thống kê mô tả trên Stata

- Phối hợp các biến số (bảng 2 chiều)

**`tab {biếnhàng} {biếncột}, co ro cell`**

  - Ví dụ: `tab gioitinh caoha`
  - Ví dụ: `tab gioitinh caoha, co`
  - Ví dụ: `tab gioitinh caoha, ro`
  - Ví dụ: `tab gioitinh caoha, co ro cell`

# Thống kê mô tả trên Stata

```
tab {biếnhàng} {biếncột}, co ro cell
```

- ✓ **co**: phần trăm theo cột
- ✓ **ro**: phần trăm theo hàng
- ✓ **cell**: phần trăm theo tổng

```
tab gioitinh caoha, co
```

| gioi tinh | cao huyet ap<br>khong cao | <br>cao ha | <br>Total |
|---|---|---|---|
| nam | 149<br>46.71 | 48<br>61.54 | 197<br>49.62 |
| nu | 170<br>53.29 | 30<br>38.46 | 200<br>50.38 |
| Total | 319<br>100.00 | 78<br>100.00 | 397<br>100.00 |

# Thống kê mô tả trên Stata

```
tab {biếnhàng} {biếncột}, co ro cell
```

- ✓ **co:** phần trăm theo cột
- ✓ ro: phần trăm theo hàng
- ✓ **cell**: phần trăm theo tổng

```
tab gioitinh caoha, ro
```

| gioi tinh | cao huyet ap | | |
|---|---|---|---|
| | khong cao | cao ha | Total |
| nam | 149 | 48 | 197 |
| | 75.63 | 24.37 | 100.00 |
| nu | 170 | 30 | 200 |
| | 85.00 | 15.00 | 100.00 |
| Total | 319 | 78 | 397 |
| | 80.35 | 19.65 | 100.00 |

# Thống kê mô tả trên Stata

```
tab {biếnhàng} {biếncột}, co ro cell
```

- ✓ **co**: phần trăm theo cột
- ✓ **ro**: phần trăm theo hàng
- ✓ **cell**: phần trăm theo tổng

```
tab gioitinh caoha, cell
```

| gioi tinh | cao huyet ap<br>khong cao | <br>cao ha | <br>Total |
|---|---|---|---|
| nam | 149<br>37.53 | 48<br>12.09 | 197<br>49.62 |
| nu | 170<br>42.82 | 30<br>7.56 | 200<br>50.38 |
| Total | 319<br>80.35 | 78<br>19.65 | 397<br>100.00 |

# Thống kê mô tả trên Stata

**`tab {biếnhàng} {biếncột}, co ro cell`**

- ✓ **co**: phần trăm theo cột
- ✓ **ro**: phần trăm theo hàng
- ✓ **cell**: phần trăm theo tổng

**`tab gioitinh caoha, co ro cell`**

| gioi tinh | cao huyet ap | | Total |
|---|---|---|---|
| | khong cao | cao ha | |
| nam | 149 | 48 | 197 |
| | 75.63 | 24.37 | 100.00 |
| | 46.71 | 61.54 | 49.62 |
| | 37.53 | 12.09 | 49.62 |
| nu | 170 | 30 | 200 |
| | 85.00 | 15.00 | 100.00 |
| | 53.29 | 38.46 | 50.38 |
| | 42.82 | 7.56 | 50.38 |
| Total | 319 | 78 | 397 |
| | 80.35 | 19.65 | 100.00 |
| | 100.00 | 100.00 | 100.00 |
| | 80.35 | 19.65 | 100.00 |

# Thống kê mô tả trên Stata

- Kiểm tra phân phối

Cách 1: **`hist {biếnsố}, norm`**

– Vídụ: `hist bmi, norm`

→ Bình thường khi có dạng hình chuông

Cách 2: **`pnorm {biếnsố}`** # Normal probability plot

– Vídụ: `pnorm bmi`

→ Bình thường khi đường in đậm trùng với đường chéo

Cách 3: **`qnorm {biếnsố}`** # Quantiles of normal distribution plot

– Vídụ: `qnorm bmi`

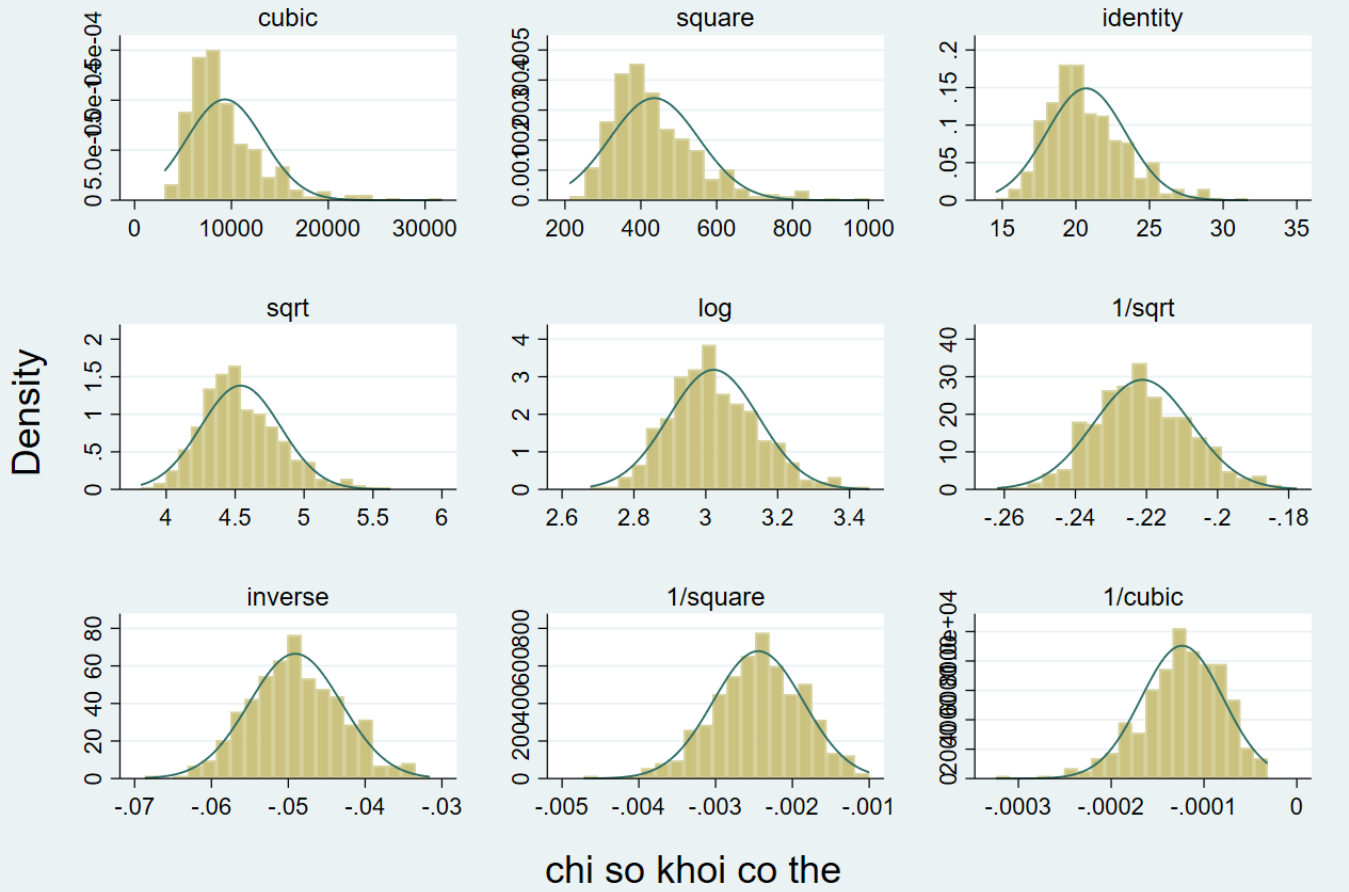
→ Bình thường khi đường in đậm trùng với đường chéo

# Thống kê mô tả trên Stata

- Kiểm tra phân phối
  - Cách 4: **`swilk {biếnsố}`** # Shapiro Wilk test
    - Ví dụ: `swilk bmi`

    → Phân phối bình thường khi p > 0,05

    → Kiểm định này rất nhạy → hạn chế dùng
- Nếu số liệu có phân phối không bình thường (lệch) thì có thể xem gợi ý cách biến đổi số liệu
  - **`gladder {biến số}`**
  - Ví dụ: `gladder bmi`


cubic
square
identity
sqrt
log
1/sqrt
inverse
1/square
1/cubic
Density
chi so khoi co the
Histograms by transformation

# Thống kê mô tả trên Stata

- Biến định lượng
  - Khi có phân phối bình thường →Trung bình (Độ lệch chuẩn)
  - Khi không có phân phối bình thường →Trung vị (Khoảng tứ vị)
  - Khoảng tin cậy 95%
  - Stata
    - ***sum {các biến số}***
      - Ví dụ: `sum hatdtb hatttb bmi`
      - Ví dụ: `sum hatdtb hatttb bmi, d`
    - ***ci mean {các biến định lượng}***
      - Ví dụ: `ci mean hatdtb hatttb bmi`

# Thống kê mô tả trên Stata

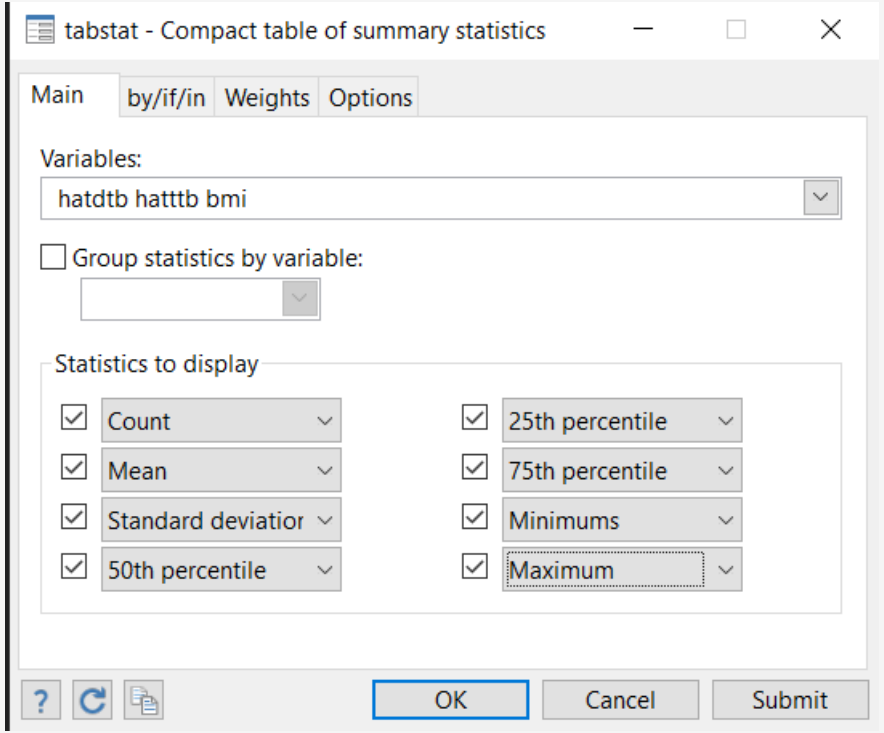
- Có thể trình bày trên cùng 1 bảng tổng hợp

```
tabstat hatdtb hatttb bmi, stat(count mean sd p50 p25 p75
min max) columns(statistics)
```

- Phân theo nhóm

```
tabstat hatdtb hatttb bmi, stat(count mean sd p50
p25 p75 min max) columns(statistics) by(caoha)
```

- Hoặc gõ **db tabstat** và chọn các biến số cùng các thông tin cần thống kê


tabstat - Compact table of summary statistics
Main
by/if/in
Weights
Options
Variables:
hatdtb hatttb bmi
Group statistics by variable:
Statistics to display
Count
Mean
Standard deviatior
50th percentile
25th percentile
75th percentile
Minimums
Maximum
OK
Cancel
Submit

# Các chỉ số thống kê mô tả (tóm tắt)

- Biến định tính
  - ✓ Tần số **`tab1 /tab`**
  - ✓ Phần trăm
  - ✓ Phần trăm cộng dồn
  - ✓ KTC 95% **`ci prop`**

    *#(mã hóa 0/1)*

- Biến định lượng
  - ❖ Có phân phối bình thường
    - ✓ Trung bình **`sum`**
    - ✓ Độ lệch chuẩn
    - ✓ Phạm vi
  - ❖ Không có phân phối bình thường
    - ✓ Trung vị **`sum , d`**
    - ✓ Khoảng tứ phân vị
    - ✓ Phạm vi
  - ❖ KTC 95% **`ci mean`**

# Nội dung đã học

- Các chỉ số thống kê mô tả
- Giới thiệu phần mềm Stata
- Thống kê mô tả trên Stata

# Thực hành