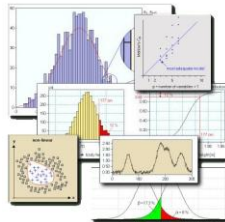


# THỐNG KÊ CƠ BẢN VÀ PHÂN TÍCH SỐ LIỆU



*PGS. TS. Hoàng Văn Minh*  
*Hà nội- tháng 12 năm 2012*

## **NỘI DUNG**

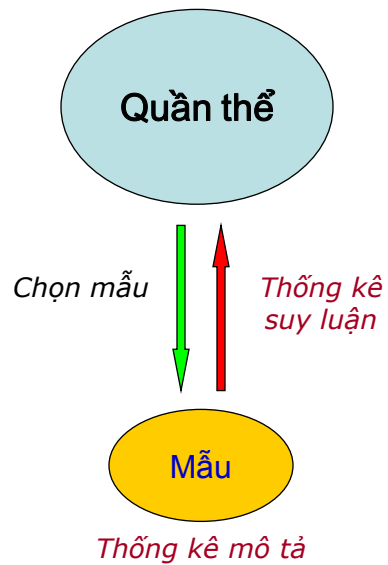
1. Khái niệm thống kê cơ bản
2. Lựa chọn trắc nghiệm thống kê
3. Thực hành thống kê với Stata

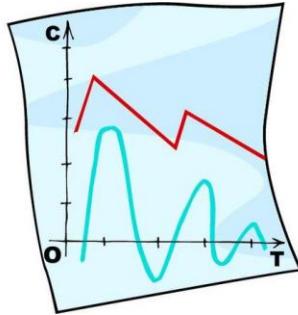
# Thống kê

"*Phân môn toán học có nhiệm vụ thu thập, phân tích, phiên giải và trình bày SỐ LIỆU*"



# Thống kê





Số liệu là đối tượng chính của thống kê

*Biến số# Hằng số*



Biến số

Định  
lượng

Định  
tính

## Số liệu ?

- Tuổi,
- Giới
- Trị số huyết áp
- Hàm lượng ure máu
- Mức độ trầm trọng của bệnh
- Các phương pháp điều trị
- Tử vong/sống sót

## THỐNG KÊ

Mô tả biến  
định lượng

Mô tả biến  
định tính

Suy luận  
biến định  
lượng

Suy luận  
biến định  
tính

## Thống kê mô tả biến định lượng

- ⊕ Đo lường độ tập trung (Location)
  - ⊕ Trung bình (mean)
  - ⊕ Trung vị (median)
  - ⊕ Mode
- ⊕ Đo lường độ phân tán (Spread )
  - ⊕ Khoảng số liệu (range)
  - ⊕ Khoảng tứ phân vị (25%-75%) (Interquartile )
  - ⊕ Độ lệch chuẩn (Standard deviation)
  - ⊕ Phương sai (Variance)

## Bài tập: tm1\_ureamau

sum urea

sum urea, d

histogram urea, normal

graph box urea

## Thống kê mô tả biến định tính

- ⊕ Tần số
- ⊕ Tỷ lệ phần trăm

## Bài tập: tm1\_ureamau

```
tab urecao
```

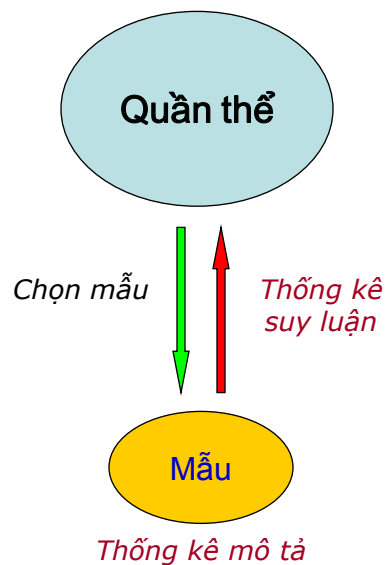
```
graph pie, over(urecao) plabel(_all percent)
```

## Bài tập: tm1\_ureamau

```
tabstat urea, by( caotuai) stat(n mean  
median sd min max) f(%9.2g)
```

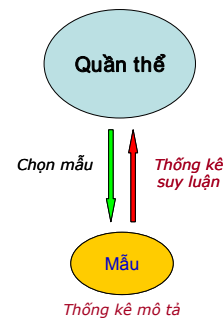
```
tab urecao caotuai, col  
graph pie, over(urecao) by(caotuai)  
plabel(_all percent)  
graph bar (mean) urecao, over(caotuai)  
lblabel(bar)
```

## Thống kê



## Thống kê suy luận

- ⊕ Ước lượng khoảng
- ⊕ Kiểm định giả thuyết



## Ước lượng khoảng-khoảng tin cậy (confidence interval)

- ⊕ Thường chọn khoảng tin cậy 95% (95%CI)
- ⊕ Khi thực hiện đo đạc 100 lần thì ít nhất 95 lần kết quả nằm trong khoảng tin cậy
- ⊕ 95% tin tưởng rằng giá trị thực của quần thể nằm trong khoảng tin cậy

$$95\%CI = \text{Trung bình} \pm 1,96 * \text{sai số chuẩn}$$



Confidence interval (%)	Multiplying factor
90	1.64
95	1.96
99	2.58
99.9	3.29

## Sai số chuẩn (standard errors)

*95%CI= Trung bình ± 1,96\*sai số chuẩn*

$$s.e. = \frac{\sigma}{\sqrt{n}}$$

$$s.e. = \sqrt{\frac{p(1-p)}{n}}$$

## Bài tập: tm1\_ureamau

ci urea

by caotuai, sort: ci urea

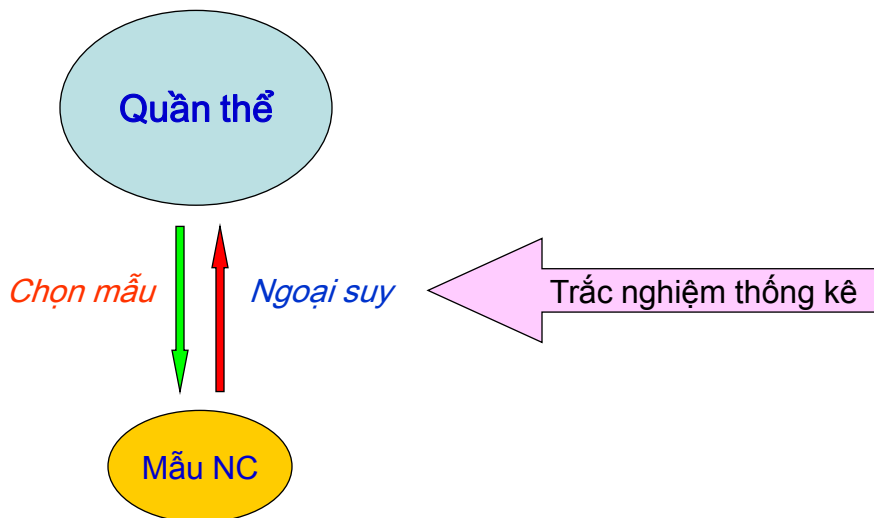
ci urecao

by caotuai, sort: ci urecao

proportion urecao

## Kiểm định giả thuyết

*sử dụng trắc nghiệm (test) thống kê để đưa ra kết luận về giả thuyết của nhà nghiên cứu là chấp nhận được hay không*



## Giả thuyết

- ⊕ Giả thuyết Ho: Không có sự khác biệt
- ⊕ Giả thuyết Ha: Có sự khác biệt



## Giả thuyết

*Tác dụng của tiêm nitrate tĩnh mạch có tác dụng giảm tỷ lệ tử vong ở bệnh nhân nhồi máu cơ tim?*

- ⊕ Giả thuyết Ho:
- ⊕ Giả thuyết Ha:



## Sai lầm

Thực tế→	$H_0$ đúng	$H_0$ sai
↓ Quyết định		
Chấp nhận $H_0$	✓	Sai lầm II ( $\beta$ )
Loại bỏ $H_0$	Sai lầm I ( $\alpha$ )	✓

## Phiên giải kết quả

*Tỷ lệ tử vong ở bệnh nhân nhồi máu cơ tim được tiêm nitrate tĩnh mạch thấp hơn tỷ lệ tử vong ở những bệnh nhân bị nhồi máu cơ tim mà không được tiêm, sự khác biệt là có ý nghĩa thống kê ( $p < 0.05$ )*



## Phiên giải kết quả

Có sự khác biệt không?  
(Nêu rõ sự khác biệt)

*Tỷ lệ tử vong ở bệnh nhân nhồi máu cơ tim được tiêm nitrate tĩnh mạch thấp hơn*



Có ý nghĩa thống kê không?

*Sự khác biệt là có ý nghĩa thống kê*



Căn cứ vào đâu?

*( $p < 0.05$ , trắc nghiệm khi bình phương)*

## Mức ý nghĩa thống kê

### Loại bỏ sai lầm loại I

$\alpha = 0.05$

$p = \text{probability} = \text{Xác suất để giả thuyết } H_0 \text{ đúng}$

- ›  $P < 0.05$  = Xác suất để giả thuyết  $H_0$  đúng là  $< 5\%$ 
  - =  $H_0$  xảy ra chỉ là may rủi
  - = Bác bỏ  $H_0$
  - = Xác suất để giả thuyết  $H_a$  đúng là  $> 95\%$
  - =  $H_a$  xảy ra là chắc chắn
  - = Chấp nhận  $H_a$
- ›  $P > 0.05 = ???$

## Mức ý nghĩa thống kê

$\alpha$	$z (1-\alpha/2)$
.01 (99)	2.576
.02 (98)	2.326
.05 (95)	1.960
.10 (90)	1.645

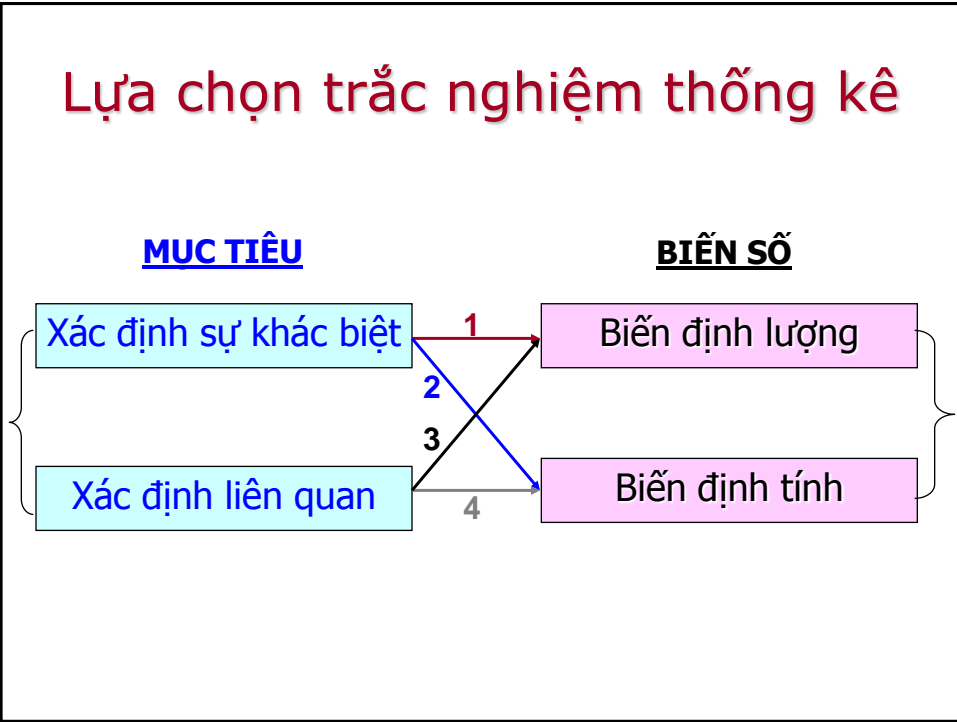
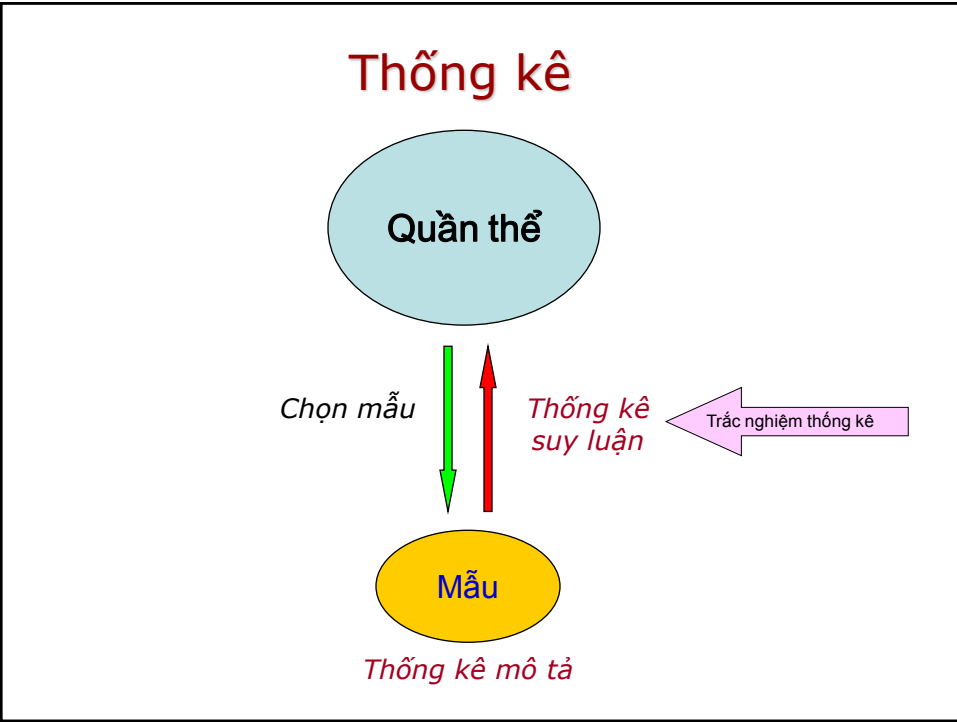
## Độ mạnh

Loại bỏ sai lầm loại II

$$1 - \beta = 80\%$$

Thường dùng trong tính toán cỡ mẫu

$1-\beta$	$z (1-\beta)$
.80	0.842
.85	1.036
.90	1.282
.95	1.645



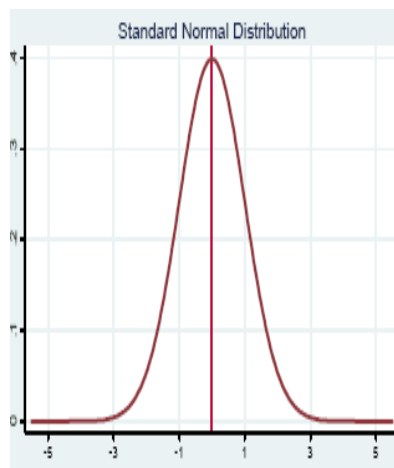
Xác định sự khác biệt  
của biến định lượng

Test tham số  
(Phân bố chuẩn)

Tes phi tham số  
(Phân bố chuẩn &  
không chuẩn)



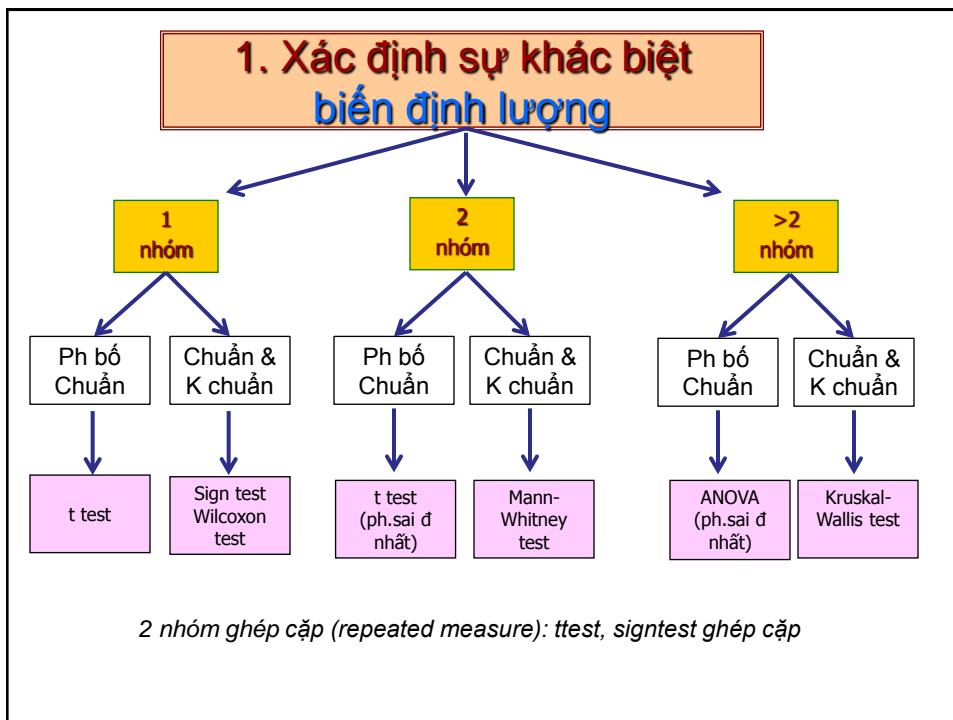
Kiểm định phân bố số liệu





# Kiểm định phân bố số liệu

- histogram urea, bin (40) normal
- sum urea, d
- sktest urea
- gladder urea



## Bài tập: tm1\_ureamau

ttest urea=5

signtest urea=5

ttest urea,by(caotuai)

oneway urea caotuai

ranksum urea,by(caotuai)

## Bài tập: tm2\_oxygen

ttest vaovien= sau6h

signtest vaovien= sau6h

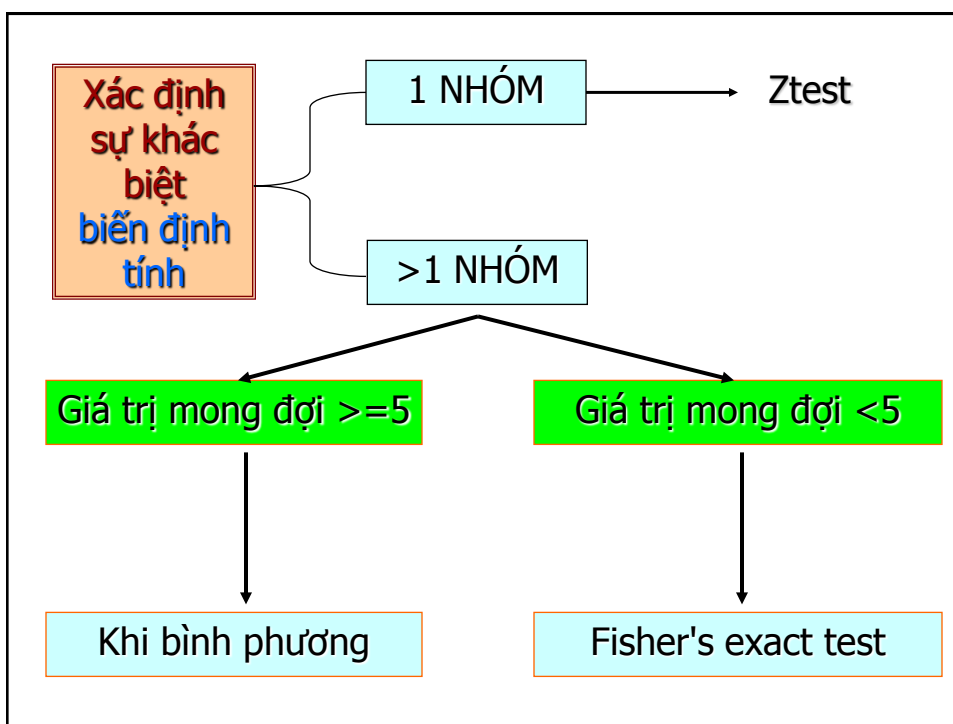
## Bài tập (*tm3\_alht.dta*)

Kiểm định sự khác biệt về áp lực huyết tương  
của các bệnh nhi nhập viện theo 3 nhóm tuổi?

*oneway alht tuoi*

*kwallis alht, by( tuoi)*

*kwallis2 alht, by( tuoi)*



## Giá trị quan sát và mong đợi

	Tốt	Không tốt	Tổng
Điều trị 1	4	5	9
Điều trị 2	3	3	6
Tổng	7	8	15

**Giá trị mong đợi = (Tổng hàng \* Tổng cột) / Tổng chung**

## Bài tập (tm4\_benhcotim)

*So sánh tỷ lệ bệnh cơ tim*

*ci benhcotim*

*prtest benhcotim=.35*

*tab benhcotim hutthuoc, expect*

*tab benhcotim hutthuoc, col chi2*

*tab benhcotim hutthuoc, col exact*

## Xác định mối liên quan

### 3. Biến định lượng:

- Hệ số tương quan ( $r$ )
  - pearson
  - spearman
- Hồi quy tuyến tính

### 4. Biến định tính

- Tỷ suất chênh ( $OR$ ), nguy cơ tương đối ( $RR$ )
- Hồi quy logistic

## Tương quan biến định lượng ( $r$ )

- Hệ số tương quan  $r$ 
  - Có giá trị từ -1 đến +1
    - Khi HSTQ  $> 0 \Rightarrow$  tương quan đồng biến
    - Khi HSTQ  $< 0 \Rightarrow$  tương quan nghịch biến
    - Càng gần 1  $\Rightarrow$  tương quan càng chặt
  - Quy ước:
    - $< 0,3$ : tương quan yếu
    - $\geq 0,3-0,5$ : tương quan TB
    - $\geq 0,5-0,7$ : tương quan chặt chẽ
    - $\geq 0,7$ : tương quan rất chặt chẽ

## Bài tập (tm5\_songcon1)

*twoway scatter tuoi thoigian*

*twoway (scatter tuoi thoigian) (lfit hatdtb tuoi)*

*corr tuoi thoigian*

*pwcorr tuoi thoigian, sig*

*spearman tuoi thoigian*

## Tương quan của 2 biến định tính

Khi muốn tìm cường độ mối liên quan giữa hai biến định tính=> có thể sử dụng:

- Tỷ suất chênh: OR
- Nguy cơ tương đối: RR

	Tốt	Không tốt	Tổng
Điều trị 1	4	5	9
Điều trị 2	3	3	6
Tổng	7	8	15

## OR và RR

- $>1$   $\Rightarrow$  Yếu tố nguy cơ
- $=1$   $\Rightarrow$  Không liên quan
- $<1$   $\Rightarrow$  Yếu tố bảo vệ

CI không chứa 1  $\Rightarrow$  Có ý nghĩa thống kê

## Bài tập

*Liên quan giữa phương pháp điều trị và nguy cơ tử vong?*

*tabodds ketqua dieutri, or*

*recode dieutri 1=0 2=1*

*cc ketqua dieutri*

*cs ketqua dieutri*

## Hiệu quả điều trị

- **RR: Relative Risk= nguy cơ tương đối**  
*Tỷ số giữa 2 tỷ lệ ở nhóm can thiệp và nhóm chứng*
- **RRR: Relative Risk Reduction = Giảm nguy cơ tương đối**  
*Mức giảm (%) biến cố ở nhóm can thiệp so với nhóm chứng*
- **ARR: Absolute Risk Reduction = Giảm nguy cơ tuyệt đối**  
*Sự khác biệt về con số tuyệt đối của tỉ lệ biến cố giữa nhóm can thiệp và nhóm chứng*
- **NNT: Number Needed to Treat = Số cá thể cần can thiệp**  
*Số cá thể cần can thiệp để tránh khỏi việc xảy ra 1 biến cố xấu*

## Nghiên cứu hiệu quả can thiệp

	Can thiệp A	Can thiệp B
Bệnh	30	50
Không bệnh	220	200
<b>Tổng</b>	<b>250</b>	<b>250</b>

Tỷ lệ bệnh can thiệp A:	$30/250=12\%$
Tỷ lệ bệnh can thiệp B:	$50/250=20\%$
Nguy cơ tương đối (RR):	$(30/250)/(50/250)=0.6$
Giảm nguy cơ tương đối (RRR):	$ 12\%-20\%  / 20\%=40\%$
Giảm nguy cơ tuyệt đối (ARR):	$ 12\%-20\% =8\%$
Số cá thể cần can thiệp (NNT):	$1/8\%=12.5$



## Bài tập

	<b>Can thiệp A</b>	<b>Can thiệp B</b>
Bệnh	25	60
Không bệnh	175	140
<b>Tổng</b>	<b>200</b>	<b>200</b>

csi 25 60 175 140