Bài giảng 14:
Phân tích 2 nhóm: biến liên tục

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Khác biệt giữa phụ nữ Việt Nam và Mĩ

Table 1 Basic characteristics of participants

Variable	US white (n = 419)	Vietnamese (n = 210)	P value
Age (years)	71.5 (8.1)	61.7 (9.6)	<0.0001
Weight (kg)	66.7 (12.9)	53.3 (7.9)	<0.0001
Height (cm)	160.8 (6.1)	148.9 (5.7)	<0.0001
BMI (kg/m^2)	25.8 (4.8)	24.1 (3.2)	<0.0001
Femoral neck BMD (g/cm^2)	0.69 (0.12)	0.63 (0.11)	<0.0001
Lumbar spine BMD (g/cm^2)	0.98 (0.19)	0.76 (0.14)	<0.0001
Whole body BMD (g/cm^2)	1.05 (0.13)	0.89 (0.11)	<0.0001
Lean mass (kg)	38.6 (5.4)	32.3 (4.1)	<0.0001
Lean mass index (kg/m^2)	14.8 (1.8)	14.6 (1.5)	0.0730
Fat mass (kg)	24.8 (8.1)	18.8 (4.9)	<0.0001
Percent body fat (%)	36.4 (6.5)	35.0 (6.2)	0.0122

Lan T Ho-Pham , et al obesity 2010

Vitamin D ở người Việt theo giới tính

- Vitamin D (25-hydroxyvitamin D)

	Men	Women
N	222	336
Mean	28.57	23.79
SD (standard deviation)	8.94	7.86

Có khác biệt giữa nam và nữ ?

Nội dung

- Vài ví dụ
- Lí thuyết
- t-test và R

Suy luận về khác biệt giữa 2 nhóm

- *Estimation (ước tính)* và *test of hypothesis (kiểm định giả thuyết)*

Giả định:

- Hai nhóm *độc lập*
- Cỡ mẫu tương đối *large*. Có thể $n_1 > 30$ và $n_2 > 30$
- Cả hai nhóm được chọn ngẫu nhiên

Ước tính: *sample* và *population*

	Sample		Population	
	Men	Women	Men	Women
N	222 (n_1)	336 (n_2)	Infinite	Infinite
Mean	28.57 (x_1)	23.79 (x_2)	μ_1 = ?	μ_2 = ?
SD (standard deviation)	8.94 (s_1)	7.86 (s_2)	σ_1 = ?	σ_2 = ?

Estimation: *sample* and *population*

	Sample		Population	
	Men	Women	Men	Women
N	222 (n_1)	336 (n_2)	Infinite	Infinite
Mean	28.57 (x_1)	23.79 (x_2)	$\mu_1 = ?$	$\mu_2 = ?$
SD (standard deviation)	8.94 (s_1)	7.86 (s_2)	$\sigma_1 = ?$	$\sigma_2 = ?$
Difference	$d = x_1 - x_1$		$\delta = \mu_1 - \mu_2$	
Status	Known		Unknown	

	Sample		Population	
	Men	Women	Men	Women
N	222 (n_1)	336 (n_2)	Infinite	Infinite
Mean	28.57 (x_1)	23.79 (x_2)	$\mu_1 = ?$	$\mu_2 = ?$
SD (standard deviation)	8.94 (s_1)	7.86 (s_2)	$\sigma_1 = ?$	$\sigma_2 = ?$
Difference	$d = x_1 - x_1$		$\delta = \mu_1 - \mu_2$	
Status	Known		Unknown	

- "*Is there real difference between men and women*" có nghĩa là $d = 0$.
- Chúng ta cần tính độ dao động mẫu của *d* (sampling variability)

Ước tính

- Chúng ta cần ước tính *d* *và* standard deviation của *d* *(*kí hiệu by s)

$$S = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Khoảng tin cậy 95% của *d:*

$$d \pm 1.96\ s$$

Test of hypothesis

Null hypothesis

$$H_o : \mu_1 = \mu_2$$

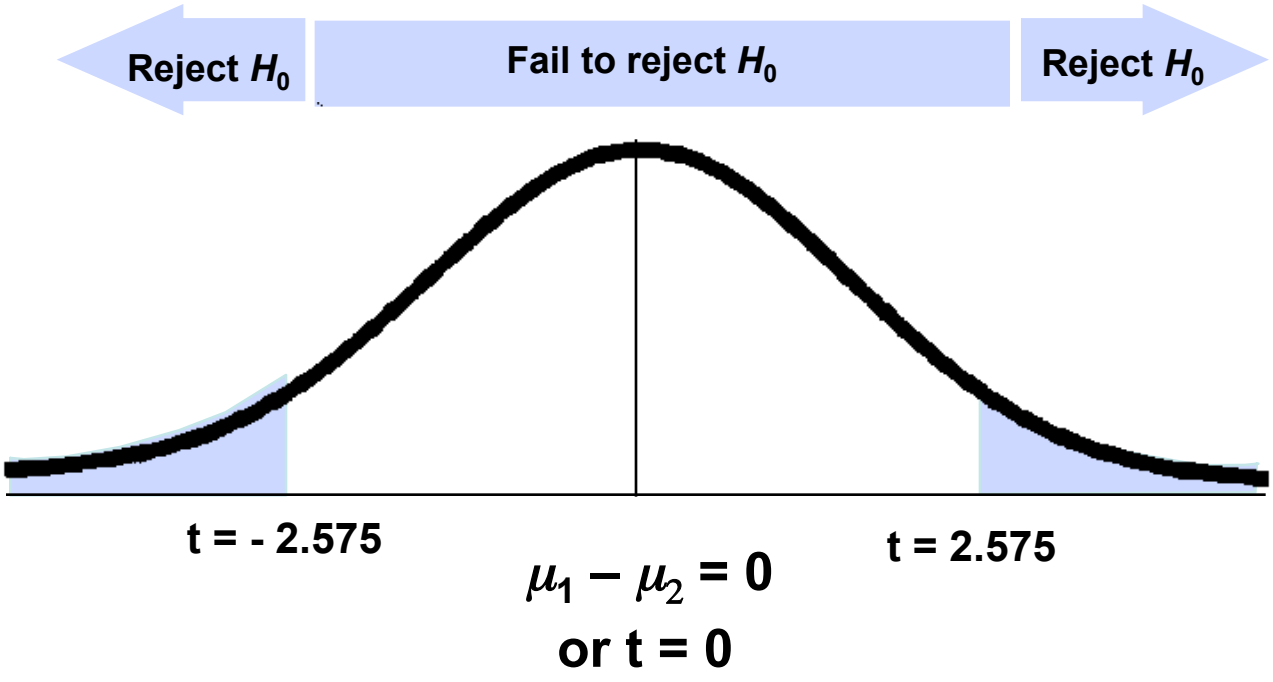
Alternative hypothesis

$$H_1 : \mu_1 \neq \mu_2$$

Câu hỏi: Nếu *Ho là thật*, xác suất mà chúng ta quan sát dữ liệu là bao nhiên? ➔ P-value

Test of hypothesis – kiểm định giả thuyết

- Đặt alpha = 0.05 hay 0.01
- Tính chỉ số thống kê (t statistic)
- So sánh chỉ số thống kê với phân bố nếu H_0 là đúng



Test statistic

$$t = \frac{\text{Difference}}{\text{SD of difference}} = \frac{\text{Signal}}{\text{Noise}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Using R

```
setwd("C:/Documents and Settings/Tuan/My Documents/
   _Current Projects/_Vietnam/Huong/Vitamin D")

vd = read.csv("vitaminD.csv", header=T, na.strings=" ")

attach(vd)

library(psych)

describe.by(vitd, sex)

t = t.test(vitd ~ sex)

print(t)
```

R outputs

```
> describe.by(vitd, sex)
group: 1
  var   n  mean   sd median trimmed  mad min   max range skew kurtosis  se
1   1 222 28.57 8.94  28.29   28.45 8.46   4 59.87 55.87 0.17     0.66 0.6
------------------------------------------------------------------------
group: 2
  var   n  mean   sd median trimmed mad min  max range skew kurtosis   se
1   1 336 23.79 7.86  24.26   23.93 7.1   4 59.6  55.6 0.07     1.22 0.43
```

R output – t-test

```
> t = t.test(vitd ~ sex)
> print(t)

t = 6.4768, df = 430.332, p-value = 2.562e-10
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 3.326365 6.224809
```

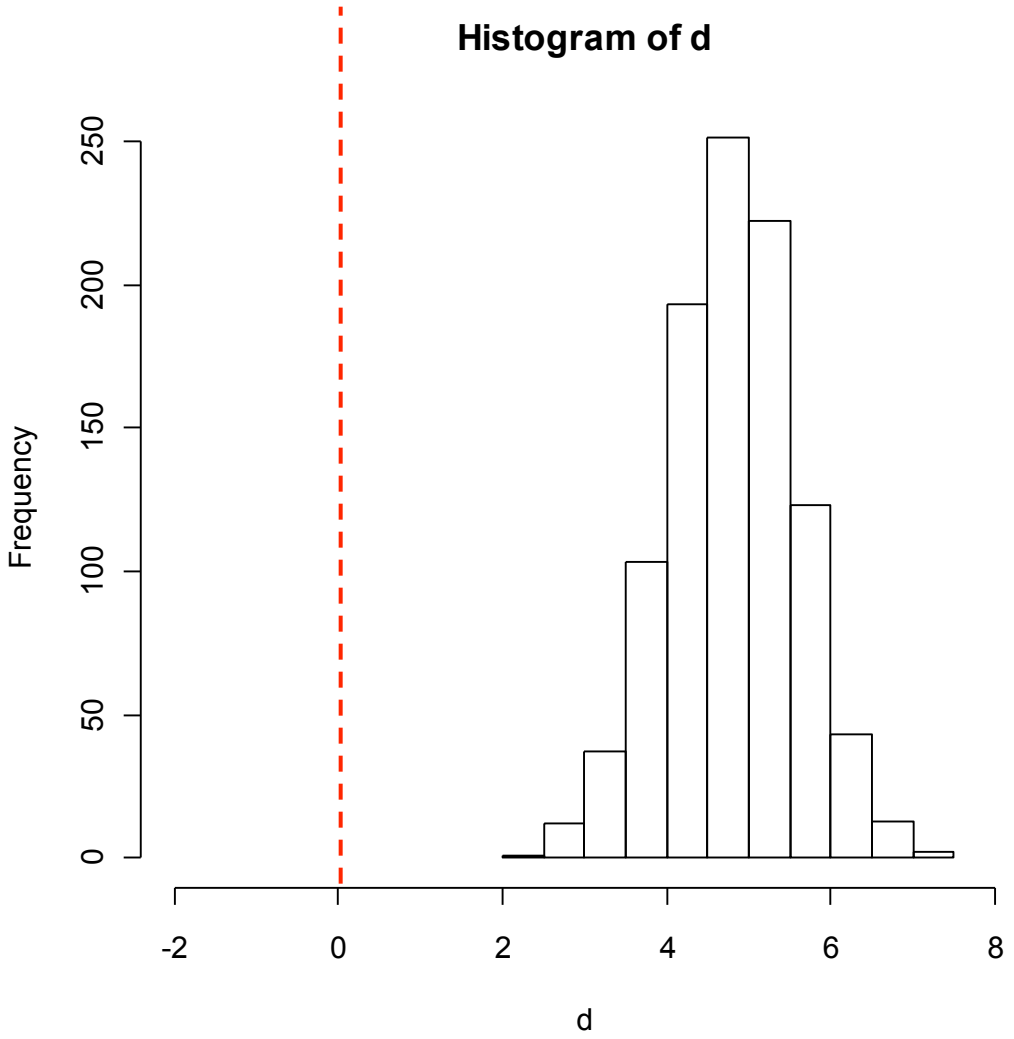
	Men	Women	Difference and 95% CI	P-value
N	222	336	336	
Mean	28.6 (8.9)	23.8 (7.9)	4.8 (3.3 – 6.2)	<0.0001

Diễn giải

	Men	Women	Difference and 95% CI	P-value
N	222	336	336	
Mean	28.6 (8.9)	23.8 (7.9)	4.8 (3.3 – 6.2)	<0.0001

25(OH)D in men was higher than that in women, with average difference being 4.8 ng/mL (95% CI: 3.3 to 6.2 ng/mL; P < 0.0001).

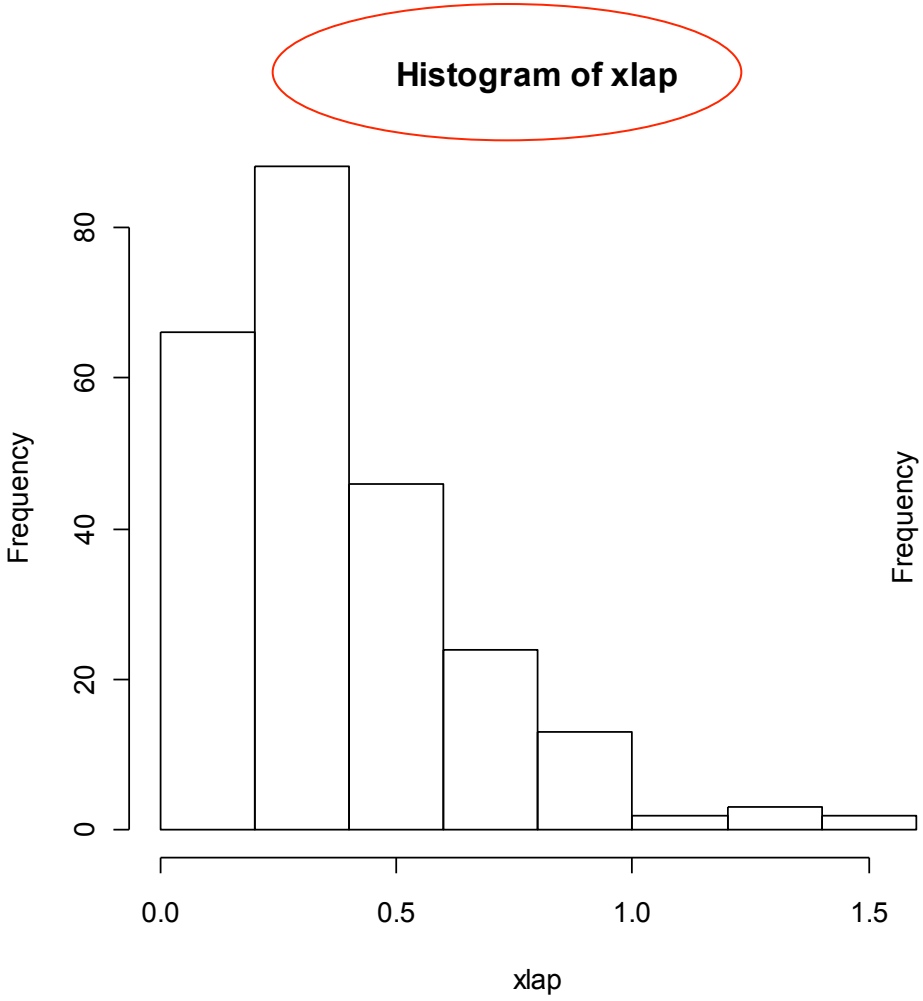
```
se = (6.2-3.3)/(2*1.96)
d = rnorm(1000, mean=4.8, sd=se)
> hist(d, xlim=c(-2, 8))
```

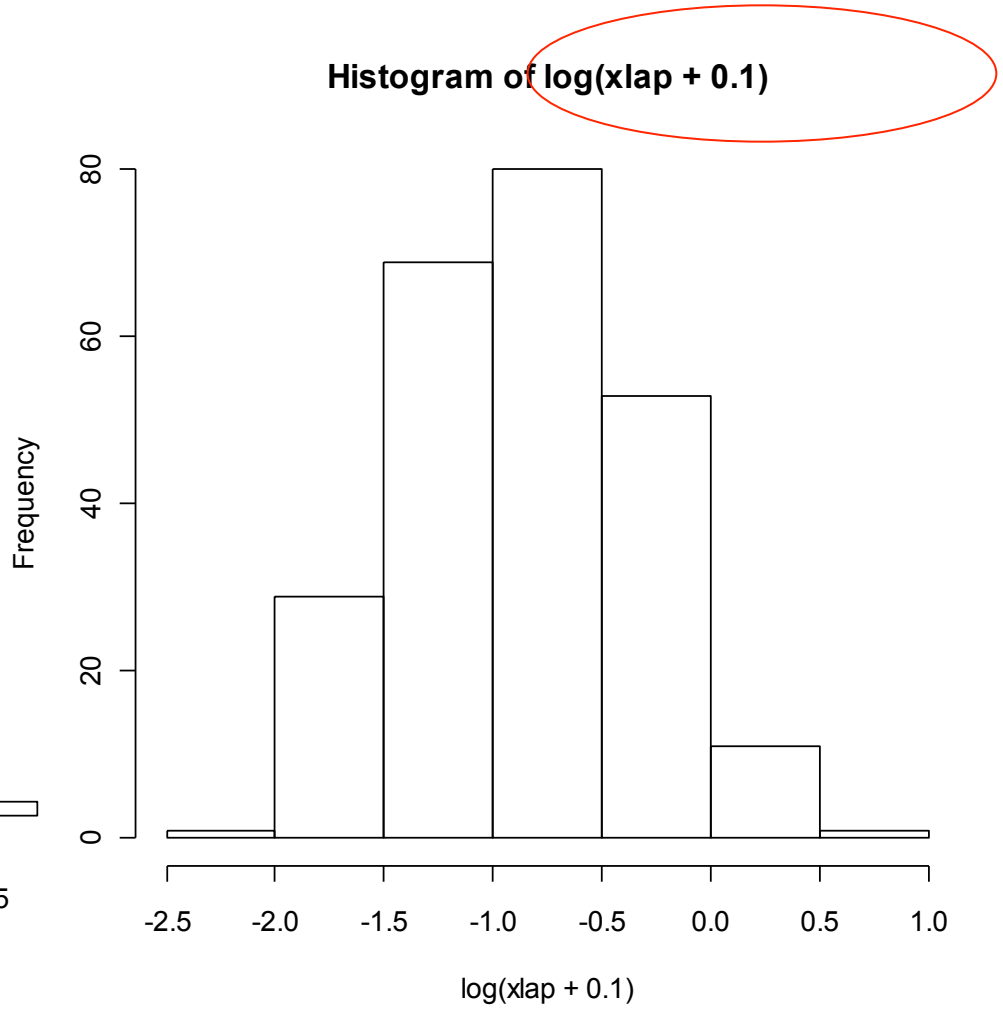


Hoán chuyển dữ liệu

```
> describe.by(xlap,  sex)

group: 1
  var   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
1   1 100 0.45 0.31   0.35    0.41 0.26 0.02 1.57  1.55  1.4     2.15 0.03
------------------------------------------------------------
group: 2
  var   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
1   1 144 0.34 0.24   0.28    0.31 0.19 0.06 1.26   1.2  1.3     1.39 0.02
```






```
> library(nortest)
> pearson.test(xlap)
Pearson chi-square normality test
data:  xlap
P = 87.877, p-value = 6.145e-12
```

```
> pearson.test(log(xlap+0.1))

Pearson chi-square normality test
data:  log(xlap + 0.1)
P = 20.7541, p-value = 0.1882
```

Phân tích dựa vào dữ liệu hoán chuyển

```
> describe.by(log(xlap+0.1),  sex)

group: 1
  var   n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
1   1 100 -0.73 0.52   -0.8   -0.73 0.54 -2.1 0.51  2.61 0.08    -0.34 0.05
------------------------------------------------------------
group: 2
  var   n  mean   sd median trimmed  mad   min  max range skew kurtosis   se
1   1 144 -0.94  0.5  -0.97   -0.97 0.52 -1.82 0.31  2.13  0.3    -0.66 0.04
```

Kết quả t test

```
> t.test(log(xlap+0.1) ~ sex)
data:  log(xlap + 0.1) by sex
t = 3.216, df = 206.284, p-value = 0.001509
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.08306888 0.34626674
sample estimates:
mean in group 1 mean in group 2
     -0.7300382      -0.9447060
```

Exp(-0.73+0.9447) = 1.234

Exp(0.083) = 1.086

Exp(0.346) = 1.413

Diễn giải

	Men	Women	Percentage difference and 95% CI	P-value
N	100	144		
Mean	0.45 (0.31)	0.34 (0.24)	23% (8.6, 41.3)	0.0015

Compared with women, beta crosslap was 23% (95% CI: 8.6 to 41.3%) **higher in men , and the difference was statistically significant (*P* = 0.001)**

Phương pháp phi tham số

- Wilcoxon's rank sum test

```
> wilcox.test(xlap ~ sex)

Wilcoxon rank sum test with continuity correction

data:  xlap by sex
W = 8890, p-value = 0.001834
alternative hypothesis: true location shift is not equal to 0
```

Phương pháp hoán vị

- Permutation và median test

```
> library(coin)
> oneway_test(xlap ~ as.factor(sex))

        Asymptotic 2-Sample Permutation Test

data:  xlap by as.factor(sex) (1, 2)
Z = 3.1073, p-value = 0.001888
alternative hypothesis: true mu is not equal to 0

> median_test(xlap ~ as.factor(sex))

        Asymptotic Median Test

data:  xlap by as.factor(sex) (1, 2)
Z = -2.8579, p-value = 0.004265
alternative hypothesis: true mu is not equal to 0
```

Tóm lược

- Statistical tests:
  - T-test: so sánh 2 nhóm, biến liên tục
  - Phương pháp phi tham số: Wilcoxon, median, and permutation tests
- Giả định: phân bố chuyển, phương sai giống nhau, độc lập
- Hoán chuyển dữ liệu, nếu cần thiết.