

Chương Trình Giảng Dạy Kinh tế Fulbright

Học kỳ Thu năm 2013

Các Phương Pháp Phân Tích Định Lượng

GỢI Ý GIẢI BÀI TẬP 5

KIỂM ĐỊNH THỐNG KÊ

Ngày Phát: Thứ Ba 29/10/2013

Ngày Nộp: Thứ Ba 05/10/2013

Bài 1: (35 điểm)

Theo một tranh luận của các học viên chương trình chính sách công Fulbright, thì giới tính của chủ hộ là một trong các yếu tố ảnh hưởng tới hành vi tiêu dùng của chính hộ đó – trong đó có tiêu chí chi tiêu cho chăm sóc sức khỏe. Hơn thế nữa, có một nhóm phát biểu rằng: “chủ hộ là “Nữ giới” sẽ quan tâm tới sức khỏe gia đình nhiều hơn nên chi tiêu cho chăm sóc sức khỏe nhiều hơn”. Để kiểm chứng điều này, anh T, học viên MPP5, đã tiến hành một cuộc điều tra dựa trên một mẫu ngẫu nhiên bao gồm 31 hộ có chủ hộ là “Nam giới” và 33 hộ có chủ hộ là “nữ giới”. Dữ liệu điều tra về mức độ chi tiêu cho y tế của mỗi hộ thể hiện trong cột “Chi tiêu y tế” của file dữ liệu **Chitieu_Yte.xls**. Với hệ thống mã hóa quy định như sau: giới tính chủ hộ được thể hiện trong cột “Giới tính” với số 1 là giới tính nam, số 2 là giới tính nữ. Dựa trên tập dữ liệu nói trên hãy làm các công việc sau đây:

- Lập các giả thuyết hợp lý.
- Với mức ý nghĩa 5%, tính các thông số cần thiết và tiến hành kiểm định giả thuyết nêu trên.
- Bạn hãy tìm giá trị P_{value} và giải thích cho họ hiểu ý nghĩa của con số này.
- Theo bạn ở mức ý nghĩa đã chọn, phát biểu trích dẫn ở trên có đáng tin cậy không?

a. Giả thuyết hợp lý được đặt ra là:

Đặt

μ_1 : Mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nữ giới

μ_2 : Mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nam giới

Trong cuộc họp, có ý kiến cho rằng chủ hộ là nữ giới quan tâm tới sức khỏe gia đình nhiều hơn nên chi tiêu cho chăm sóc sức khỏe nhiều hơn. Đây là phát biểu của một nhóm trong một cuộc tranh luận, nên ta có thể nghi ngờ về tính chính xác của ý kiến này. Có thể lập luận rằng, chủ hộ là nữ thì hộ gia đình có thể có thu nhập ít hơn so với hộ có chủ hộ là nam, nên chi tiêu cho y tế có thể cũng sẽ thấp hơn.¹

¹ Một cách lập luận khác có thể là: Chi tiêu y tế là chi tiêu mà dù chủ hộ là nam hay nữ thì cũng đều phải chi và không thể tiết kiệm được. Bởi vậy ta nghi ngờ nhận định trên.

Do vậy, ta lập các giả thuyết như sau:²

$H_0: \mu_1 - \mu_2 \geq 0$ (Mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nữ giới nhiều hơn so với nam giới)

$H_a: \mu_1 - \mu_2 < 0$ (Mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nữ giới ít hơn so với nam giới)

b) Với mức ý nghĩa $\alpha=5\%$, tính các thông số cần thiết và tiến hành kiểm định giả thuyết trên:

Từ file dữ liệu **Chitieuyte.xls**, dùng lệnh filter ở cột Giới tính, ta lọc và tách riêng dữ liệu cho Nam và Nữ. Ta được hai mảng dữ liệu, một là dữ liệu về Chi tiêu y tế của 31 hộ gia đình có chủ hộ là Nam giới, hai là dữ liệu về Chi tiêu y tế của 34 hộ gia đình có chủ hộ là Nữ giới.

Đối với mỗi mảng dữ liệu, ta dùng công cụ **Tools\Data Analysis\Descriptive Statistics** trong **Excel 2007** để tính toán chỉ tiêu thống kê cho từng nhóm, và được kết quả như sau:

<i>Chi tiêu y tế - Chủ hộ là nữ giới (1000đ/năm)</i>		<i>Chi tiêu y tế - Chủ hộ là nam giới (1000đ/năm)</i>	
Mean	1150,588235	Mean	1140,806452
Standard Error	303,3543257	Standard Error	371,1134175
Median	450	Median	520
Mode	0	Mode	0
Standard Deviation	1768,84448	Standard Deviation	2066,27206
Sample Variance	3128810,795	Sample Variance	4269480,228
Kurtosis	9,951713345	Kurtosis	15,13449834
Skewness	2,845970015	Skewness	3,711607338
Range	8750	Range	10530
Minimum	0	Minimum	0
Maximum	8750	Maximum	10530
Sum	39120	Sum	35365
Count	34	Count	31

- Đặt tên các biến số cho hai mảng dữ liệu:

$\bar{X}_1 = 1150,588235$ (1000đ/năm) (Mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nữ giới của 34 hộ gia đình)

$n_1 = 34$ (Số hộ gia đình có chủ hộ là nữ)

$\bar{X}_2 = 1140,806452$ (1000đ/năm) (Mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nam giới của 31 hộ gia đình)

$n_2 = 31$ (Số hộ gia đình có chủ hộ là nam)

$S_1 = 1768,84448$ (1000đ/năm) (Độ lệch chuẩn mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nữ giới)

² Lưu ý: Việc lập giả thuyết như thế nào là tùy thuộc quan điểm và lập luận của mỗi người. Tuy nhiên khi đưa ra giả thuyết thì cần có lập luận giải thích.

$S_2 = 2066,27206$ (1000đ/năm) (Độ lệch chuẩn mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nam giới)

- Kiểm định thống kê một phía cho mẫu lớn đối với $(\mu_1 - \mu_2 > 0)$

$$H_0: \mu_1 - \mu_2 \geq D_0 = 0$$

$$H_a: \mu_1 - \mu_2 < D_0 = 0$$

Trị kiểm định:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(1150,588235 - 1140,806452) - 0}{\sqrt{\frac{1768,84448^2}{34} + \frac{2066,27206^2}{31}}} = 0,020407567$$

Vậy ta có trị kiểm định $t = 0,020407567$, bậc tự do $n_1 + n_2 - 2 = 34 + 31 - 2 = 63$

Với mức ý nghĩa $\alpha = 5\%$, tra bảng phân phối student hoặc sử dụng lệnh **inv**, tính ra được giá trị tới hạn $t_\alpha = 1,669402222$ (có thể lấy xấp xỉ với bậc tự do 60 với giá trị $t_\alpha = 1,671$)

Do $-t_\alpha < t \Rightarrow$ Không bác bỏ giả thuyết H_0

Điều đó chứng tỏ, với mức ý nghĩa là 5%, không có đủ cơ sở để cho rằng mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nữ giới không cao hơn so với nam giới.

Chú ý:

Kiểm định này có thể thực hiện trên Stata và cho kết quả tương tự:

```
. ttest nu == nam, unpaired unequal
```

Two-sample t test with unequal variances

variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
nu	34	1150.588	303.3543	1768.844	533.4092	1767.767
nam	31	1140.806	371.1134	2066.272	382.8917	1898.721
combined	65	1145.923	235.8173	1901.22	674.8239	1617.022
diff		9.781784	479.3214		-949.2089	968.7725

```
diff = mean(nu) - mean(nam) t = 0.0204
Ho: diff = 0 Satterthwaite's degrees of freedom = 59.3825
```

```
Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.5081 Pr(|T| > |t|) = 0.9838 Pr(T > t) = 0.4919
```

c. Bạn hãy tìm giá trị Pvalue và giải thích cho họ hiểu ý nghĩa của con số này

Ta có $P_{value} = P(t < 0,020407567) = 1 - P(t > 0,020407567) = 1 - 0,491891373 = 50,81\%$

Vì $P_{value} > \alpha \Rightarrow$ Không bác bỏ giả thuyết H_0

Giải thích cho họ về ý nghĩa của P_{value} như sau: Xác suất sai lầm lớn nhất khi cho rằng mức chi tiêu cho y tế trung bình của hộ gia đình có chủ hộ là nữ giới ít hơn nam giới là 50,81%.

d. Theo bạn ở mức ý nghĩa đã chọn, phát biểu trích dẫn ở trên có đáng tin cậy không?

Vì $P_{value} > \alpha$ nên ta không bác bỏ giả thuyết H_0 ; một cách chắc chắn thì chỉ có thể kết luận được rằng, chưa có đủ cơ sở để cho rằng phát biểu trên là không đáng tin cậy.

Bài 2: (25 điểm)

Theo nhận định của hiệp hội dệt may Việt Nam, hiện nay 80% sản phẩm dệt may ở thành phố (T.p) Hồ Chí Minh được đưa đến tay người tiêu dùng thông qua kênh phân phối là các chợ truyền thống. Phần còn lại (20%) thông qua các kênh phân phối như: siêu thị, cửa hàng thời trang cao cấp hay các trung tâm giới thiệu sản phẩm. Hỏi ngẫu nhiên 60 người tiêu dùng ở T.p thì có 43 người trả lời là mua sản phẩm dệt may ở các chợ truyền thống. Với mức ý nghĩa 5%, kiểm định xem nhận định trên có đáng tin cậy không?

Tóm tắt:

$$p_0 = 0,8$$

$$n = 60$$

$$\hat{p} = 43/60 = 0,7166$$

$$\alpha = 5\%$$

Ta tiến hành kiểm định hai phía cho Mẫu lớn đối với một tỷ lệ tổng thể:

Trong trường hợp này, ta nghi ngờ nhận định của hiệp hội dệt may nên đặt các giả thuyết:

Giả thuyết H_0 : $p = p_0$ (80% sản phẩm dệt may ở thành phố HCM được đưa đến tay người tiêu dùng thông qua kênh phân phối là các chợ truyền thống, tức là báo cáo trên đáng tin cậy)

Giả thuyết thay thế H_a : $p \neq p_0$

$$\text{Trị thống kê kiểm định } z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0,7166 - 0,8}{\sqrt{\frac{0,8 \cdot (1 - 0,8)}{60}}} = -1,613743061$$

Ta có $z_{\frac{\alpha}{2}} = z_{\frac{0,05}{2}} = z_{0,025}$. Tra bảng phân phối chuẩn hóa ta có $z_{\frac{\alpha}{2}} = 1,96$

Vì $-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}$ => không bác bỏ giả thuyết H_0 .

Vậy mức ý nghĩa là 5% thì không đủ cơ sở để cho rằng việc 80% sản phẩm dệt may ở thành phố HCM được đưa đến tay người tiêu dùng thông qua kênh phân phối là các chợ truyền thống là không đáng tin cậy.

Bài 3: (30 điểm)

Một cửa hàng thực phẩm, trong một khu dân cư tương đối cô lập, nhận thấy thời gian vừa qua trung bình một khách hàng mua 280 nghìn đồng thực phẩm/ngày. Để đánh giá tác động của suy thoái kinh tế tới sức mua của khách hàng, ông chủ cửa hàng quan sát ngẫu nhiên 15 khách hàng có được bảng dữ liệu như sau:

STT	Giá trị mua (nghìn đồng)
1	257
2	142
3	158
4	430
5	100
6	19
7	56
8	500
9	350
10	200
11	800
12	90
13	100
14	230
15	190

Thông qua số liệu nói trên ông chủ cửa hàng cảm thấy dường như suy thoái kinh tế đang làm giảm sức mua của khách tại cửa hàng của mình. Với mức ý nghĩa là 5%, thử xem có phải sức mua của khách hàng hiện nay thực sự giảm sút?

Các giá trị thống kê mô tả của mẫu trên:

<i>Giá trị mua (nghìn đồng)</i>	
Mean	241,4666667
Standard Error	53,21035226
Median	190
Mode	100
Standard Deviation	206,0828081
Sample Variance	42470,12381
Kurtosis	2,829183303
Skewness	1,612140391
Range	781

Minimum	19
Maximum	800
Sum	3622
Count	15

Tóm tắt bài tập:

$$\mu_0 = 280 \text{ nghìn đồng/ngày}$$

$n=15$ (mẫu nhỏ => sử dụng phân phối t)

$$\bar{x} = 241,4666667 \text{ nghìn đồng/ngày}$$

$$s^2 = 42470,12381 \text{ (nghìn đồng)}^2$$

$$s = 206,0828081 \text{ nghìn đồng}$$

$$\alpha = 5\%$$

Ta tiến hành kiểm định cho mẫu nhỏ về một giả thuyết liên quan đến trung bình tổng thể. Ở đây dự đoán rằng sức mua của khách hàng có thể đã thực sự giảm sút.

Giả thuyết $H_0: \mu = \mu_0$ (sức mua của khách hàng không giảm sút)

Giả thuyết thay thế $H_a: \mu < \mu_0$ (sức mua của khách hàng thực sự giảm sút)

$$\text{Trị kiểm định } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{241,4666667 - 280}{206,0828081/\sqrt{15}} = -0,72416986$$

Với mức ý nghĩa là $\alpha = 5\%$, tra bảng phân phối student với $\alpha = 5\%$; số bậc tự do là $n-1=14$, ta có: $t_\alpha=1,761$

$\Rightarrow t > -t_\alpha \Rightarrow$ không bác bỏ giả thuyết H_0 . Điều này cho thấy, với mức ý nghĩa là 5% thì không có đủ cơ sở để cho rằng sức mua của khách hàng thực sự giảm sút.

Chú ý: Có thể thực hiện kiểm định này trên stata và ra kết quả tương tự như sau:

```
. ttest giatrimua == 280
```

```
One-sample t test
```

variable	obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
giatri~a	15	241.4667	53.21035	206.0828	127.3418	355.5915

```
mean = mean(giatrimua)                                t = -0.7242
Ho: mean = 280                                         degrees of freedom = 14
```

```
Ha: mean < 280
Pr(T < t) = 0.2404
```

```
Ha: mean != 280
Pr(|T| > |t|) = 0.4809
```

```
Ha: mean > 280
Pr(T > t) = 0.7596
```

Bài 4: (10 điểm)

Theo văn bản Số: 1346 /BC-UBXH12 gửi cho Quốc hội với nội dung “*Báo cáo kết quả giám sát tình hình thực hiện bình đẳng giới và việc triển khai thi hành luật bình đẳng giới*” phó chủ nhiệm

Ủy ban Các vấn đề Xã hội đã phát biểu những ý sau đây:

“Vị thế việc làm của lao động nữ cũng có sự thay đổi tích cực. Trong 10 năm từ 1997 đến 2007, nhóm lao động làm công ăn lương tăng rất mạnh trong cơ cấu phân bố lao động, từ 18,6% (1997) lên tới 30% (2007), trong đó lao động nam chiếm 59,8% và lao động nữ chiếm 40,2% (2007).”¹

“Ở bậc đại học và cao đẳng, tỷ lệ sinh viên, học sinh là nữ đang có xu hướng tăng dần đều qua các năm ..., trong đó năm học 2006-2007 tỷ lệ học sinh, sinh viên nữ còn cao hơn tỷ lệ nam học sinh, sinh viên (nữ 53,32%, nam 46,68)”²

“Nhìn chung, việc quan tâm chăm sóc sức khỏe luôn tăng đều ở cả nam, nữ và không có sự khác biệt giới trong việc khám bệnh, chữa bệnh nội trú, ngoại trú giữa nam và nữ (xem bảng 12 của Phụ lục).”³

Khai thác bộ dữ liệu điều tra mức sống dân cư VHLSS 2008, các bạn hãy kiểm định về tính chính xác của các con số có liên quan tới nữ giới (của báo cáo ở năm 2007) cũng như phát biểu về việc quan tâm chăm sóc sức khỏe. (Mức ý nghĩa 5%)

Bài 4:

Khai thác bộ dữ liệu điều tra mức sống dân cư VHLSS 2008, các bạn các bạn hãy kiểm định về tính chính xác của các con số có liên quan tới nữ giới (của báo cáo năm 2007) cũng như phát biểu về việc quan tâm chăm sóc sức khỏe (mức ý nghĩa 5%)

a) *Vị thế việc làm của lao động nữ cũng có sự thay đổi tích cực, lao động nam chiếm 59,8% và lao động nữ chiếm 40,2%.*

- Trong Stata 11, ta mở file **muc123a.dta** bằng cách sử dụng lệnh:

```
. use "E:\z.VHLSS2008\muc123a.dta", clear  
. sort tỉnh huyện xa diaban hoso matv  
. save "E:\z.VHLSS2008\muc123a_sorted.dta", replace  
file E:\z.VHLSS2008\muc123a_sorted.dta saved
```

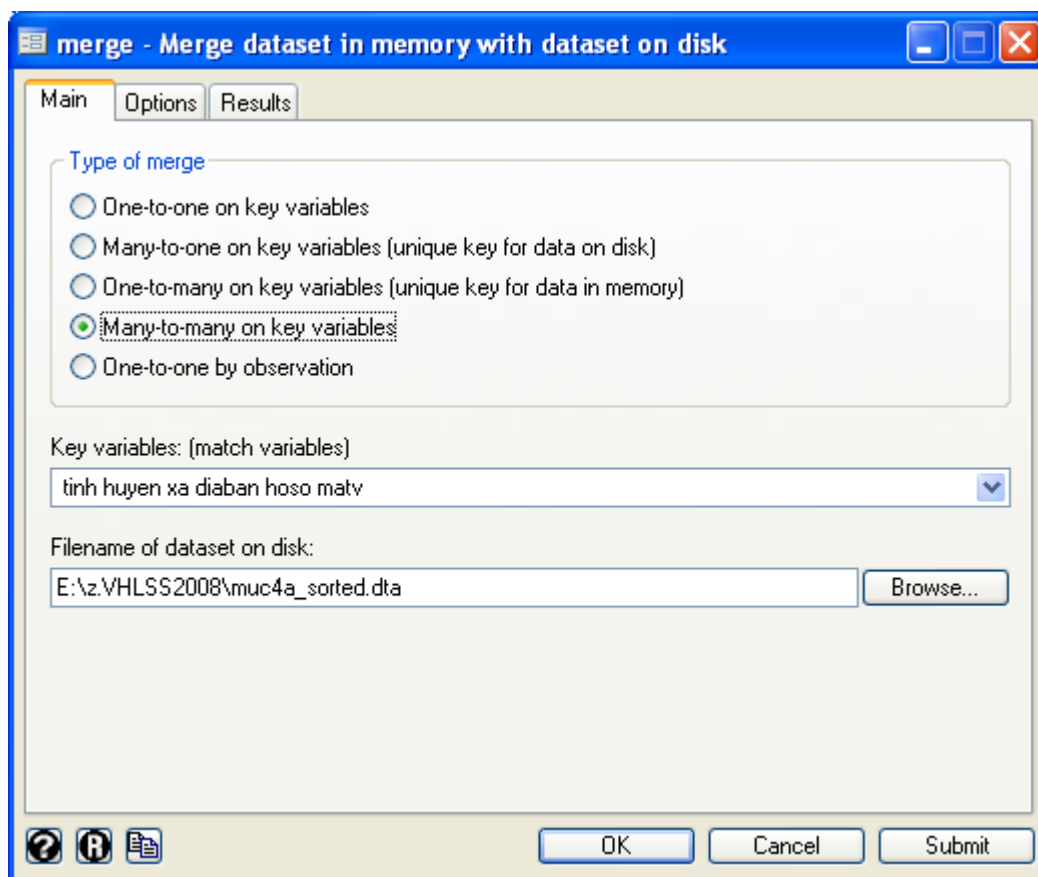
Sử dụng lệnh **sort** tương tự cho các tập tin **muc4a.dta**; **muc3b.dta**. Như vậy ta được các file **muc3b_sorted.dta**; **muc4a_sorted.dta** và **muc123a_sorted.dta**.

Thông tin chứa dữ liệu về lao động làm công ăn lương chứa trong file **muc4a**, vậy ta cần merge dữ liệu vào file master **muc123a** bằng cách:

```
use "E:\z.VHLSS2008\muc123a_sorted.dta", clear
```

- Trong Stata, chọn **Data -> Combine datasets -> Merge two datasets**.

Sau đó trong cửa sổ **Merge – Merge Dataset in memory with dataset on disk**, trong thẻ **Main** chọn **Many to many on key variables**; trong mục **Filename of dataset on disk** chọn tới file **muc4a_sorted.dta**



- Ta được kết quả như sau:

```
. merge m:m tinh huyen xa diaban hoso matv using "E:\z.VHLSS2008\muc4a_sorted.dta"
```

Result	# of obs.	
not matched	3,099	
from master	3,099	(_merge==1)
from using	0	(_merge==2)
matched	35,154	(_merge==3)

```
. keep if _merge==3
```

(3099 observations deleted)

- Để biết dữ liệu về tỉ lệ lao động Nam, Nữ thuộc nhóm làm công ăn lương, ta sử dụng lệnh **tab m1ac2 m4ac1a**, ta được kết quả như sau:

```
. tab m1ac2 m4ac1a
```

2. Giới tính	1A. Loại nghề làm công ăn lương		Total
	Công cấp	Khác	
Nam	5,736	11,466	17,202
Nữ	3,672	14,280	17,952
Total	9,408	25,746	35,154

- Kiểm định tỷ lệ làm công ăn lương = 30%

- Ta tiến hành Kiểm định cho Mẫu Lớn đối với một Tỷ lệ Tổng thể như sau:

Gọi: $p_0 = 30\%$ là tỷ lệ lao động làm công ăn lương (cả nam và nữ)

$$\hat{p} = \frac{9408}{35154} = 0,267622461 \text{ là tỷ lệ lao động nữ làm công ăn lương trong mẫu khảo sát}$$

$n = 35154$ là số lượng mẫu khảo sát

Giả thuyết $H_0: \hat{p} = p_0$

Giả thuyết thay thế: $H_a: p < p_0$

Trị thống kê kiểm định:

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0,267622461 - 0,3}{\sqrt{\frac{0,3(1-0,3)}{35154}}} = -13,24712272$$

Vùng bác bỏ:

Ta có $z_{\alpha} = z_{0,05} = 1,65$

Do $z < -z_{\alpha}$, vậy ta có đủ cơ sở bác bỏ H_0 .

Do đó, với mức ý nghĩa $\alpha = 5\%$, có đủ cơ sở để cho rằng tỷ lệ lao động có công ăn việc làm nhỏ hơn 30%, tức là tỉ lệ trong báo cáo đang cao hơn so với thực tế.

Chú ý: Có thể kiểm định thông qua Stata bằng cách sử dụng cách lệnh như sau và cho kết quả tương tự:

`. gen LCAL= (m4a<1a==1)`

`. tab LCAL`

LCAL	Freq.	Percent	Cum.
0	25,746	73.24	73.24
1	9,408	26.76	100.00
Total	35,154	100.00	

`. prtest LCAL = 0.3`

One-sample test of proportion LCAL: Number of obs = 35154

variable	Mean	Std. Err.	[95% Conf. Interval]	
LCAL	.2676225	.0023612	.2629945	.2722504

$p = \text{proportion(LCAL)} \quad z = -13.2471$
 $H_0: p = 0.3$

$H_a: p < 0.3 \quad \text{Pr}(Z < z) = 0.0000$

 $H_a: p \neq 0.3 \quad \text{Pr}(|Z| > |z|) = 0.0000$

 $H_a: p > 0.3 \quad \text{Pr}(Z > z) = 1.0000$

• **Kiểm định tỷ lệ việc lao động nữ chiếm 40,2%**

- Ta tiến hành Kiểm định cho Mẫu Lớn đối với một Tỷ lệ Tổng thể như sau:

Gọi: $p_0 = 40,2\%$ là tỷ lệ lao động nữ làm công ăn lương

$$\hat{p} = \frac{3672}{9408} = 0,3903 \text{ là tỷ lệ lao động nữ làm công ăn lương trong mẫu khảo sát}$$

$n = 9408$ là số lượng mẫu khảo sát

Giả thuyết $H_0: \hat{p} = p_0$

Giả thuyết thay thế: $H_a: p < p_0$

Trị thống kê kiểm định:

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0,3903 - 0,402}{\sqrt{\frac{0,402(1 - 0,402)}{9408}}} = -2,313361895$$

Vùng bác bỏ:

Ta có $z_{\alpha} = z_{0,05} = 1,65$

Do $z < -z_{\alpha}$, vậy ta có đủ cơ sở bác bỏ H_0 .

Do đó, với mức ý nghĩa $\alpha = 5\%$, có đủ cơ sở để cho rằng tỷ lệ lao động nữ nhỏ hơn 40,2%, tức là tỉ lệ trong báo cáo đang cao hơn so với thực tế.

Chú ý: Có thể kiểm định thông qua Stata bằng cách sử dụng cách lệnh như sau và cho kết quả tương tự:

```
gen nuLCAL=1 if (m4ac1a==1) & (m1ac2==2)
```

```
replace nuLCAL=0 if (m4ac1a==1) & (m1ac2==1)
```

```
tab nuLCAL
```

nuLCAL	Freq.	Percent	Cum.
0	5,736	60.97	60.97
1	3,672	39.03	100.00
Total	9,408	100.00	

```
prtest nuLCAL = 0.402
```

One-sample test of proportion nuLCAL: Number of obs = 9408

variable	Mean	Std. Err.	[95% Conf. Interval]	
nuLCAL	.3903061	.0050293	.3804488	.4001634

$p = \text{proportion}(\text{nuLCAL})$ $z = -2.3134$
 $H_0: p = 0.402$

$H_a: p < 0.402$ $H_a: p \neq 0.402$ $H_a: p > 0.402$
 $\Pr(Z < z) = 0.0104$ $\Pr(|Z| > |z|) = 0.0207$ $\Pr(Z > z) = 0.9896$

b) Ở bậc đại học và cao đẳng, tỷ lệ sinh viên, học sinh là nữ đang có xu hướng tăng dần đều qua các năm..., trong đó năm học 2006-2007 tỷ lệ học sinh, sinh viên nữ còn cao hơn tỷ lệ nam học sinh, sinh viên (nữ 53,32%; nam 46,68%)³

- Để biết dữ liệu về tỉ lệ Nam, Nữ thuộc đang học ở các bậc học, ta sử dụng lệnh **tab m1ac2 m2ac8**, ta được kết quả như sau:

³ Học viên có thể kiểm định thêm về việc tỷ lệ học sinh, sinh viên nữ cao hơn sinh viên nam.

2. Giới tính	Như trí,	Tiểu học	THCS	8. HỒ/cẤP/bẾc Ờang hẶc THPT	Ờang hẶc S- cẤP ng	Trung cẤP	TH CN	Cao Ờang	Total
Nam	577	1,473	1,539	1,009	46	94	120	66	5,453
Nữ	507	1,377	1,493	1,012	22	27	135	32	5,105
Total	1,084	2,850	3,032	2,021	68	121	255	98	10,558

2. Giới tính	Cao Ờang	8. HỒ/cẤP/bẾc Ờang hẶc S- cẤP ng	Ờang hẶc S- cẤP ng	Ờang hẶc S- cẤP ng	Ờang hẶc S- cẤP ng	Total
Nam	99	419	10	0	1	5,453
Nữ	147	345	7	1	0	5,105
Total	246	764	17	1	1	10,558

- Ta tiến hành Kiểm định cho Mẫu Lớn đối với một Tỷ lệ Tổng thể như sau:

Gọi: $p_0 = 53,32\%$ là tỷ lệ nữ đang học ở bậc đại học và cao đẳng

$$\hat{p} = \frac{32 + 147 + 345}{98 + 246 + 764} = \frac{524}{1108} = 0,473 \text{ là tỷ lệ nữ đang học ở bậc đại học và cao đẳng trong mẫu khảo sát}$$

$n = 1108$ là số lượng mẫu khảo sát

Giả thuyết $H_0: \hat{p} = p_0$

Giả thuyết thay thế: $H_a: \hat{p} < p_0$ (theo nhận định thì nữ giới thường không được ưu tiên như nam giới trong việc học hành, do đó tỷ lệ sinh viên học đại học và cao đẳng là nữ thường thấp hơn nam giới, do đó ta đưa ra giả thuyết thay thế là tỷ lệ sinh viên nữ thấp hơn con số được đưa ra)

Trị thống kê kiểm định:

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0,473 - 0,5332}{\sqrt{\frac{0,5332(1 - 0,5332)}{1108}}} = -4,0165$$

Vùng bác bỏ:

Ta có $z_{\alpha} = z_{0,05} = 1,65$

do $z < -z_{\alpha}$, vậy ta bác bỏ H_0 .

Do đó, với mức ý nghĩa $\alpha = 5\%$, có đủ cơ sở để cho rằng tỷ lệ sinh viên nữ ở bậc học đại học và cao học phải nhỏ hơn con số đưa ra là 53,32%. Tức là tỉ lệ này trong báo cáo đang cao hơn so với thực tế.

Chú ý: Có thể kiểm định thông qua Stata bằng cách sử dụng các lệnh như sau và cho kết quả tương tự:

```
. gen NuCDDH = 1 if m1ac2==2 & ( m2ac8<=9 & m2ac8>=7)
```

(34630 missing values generated)

```
. replace NuCDDH = 0 if m1ac2==1 & ( m2ac8<=9 & m2ac8>=7)
```

(584 real changes made)

```
. tab NuCDDH
```

. tab NUCDDH

NUCDDH	Freq.	Percent	Cum.
0	584	52.71	52.71
1	524	47.29	100.00
Total	1,108	100.00	

. prtest NuCDDH = 0.5332

. prtest NUCDDH = 0.5332

One-sample test of proportion NUCDDH: Number of obs = 1108

variable	Mean	Std. Err.	[95% Conf. Interval]	
NUCDDH	.4729242	.014999	.4435267	.5023217

p = proportion(NUCDDH) z = -4.0216
 Ho: p = 0.5332

Ha: p < 0.5332 Ha: p != 0.5332 Ha: p > 0.5332
 Pr(Z < z) = 0.0000 Pr(|Z| > |z|) = 0.0001 Pr(Z > z) = 1.0000

c) Nhìn chung, việc quan tâm chăm sóc sức khỏe luôn tăng đều ở cả nam và nữ, không có sự khác biệt giới trong việc khám bệnh, chữa bệnh nội trú, ngoại trú giữa nam và nữ.

Tương tự như câu (a), ta cần nối dữ liệu các file trong Stata với nhau:

- Trong Stata 11, ta mở file **muc123a.dta** bằng cách sử dụng lệnh:

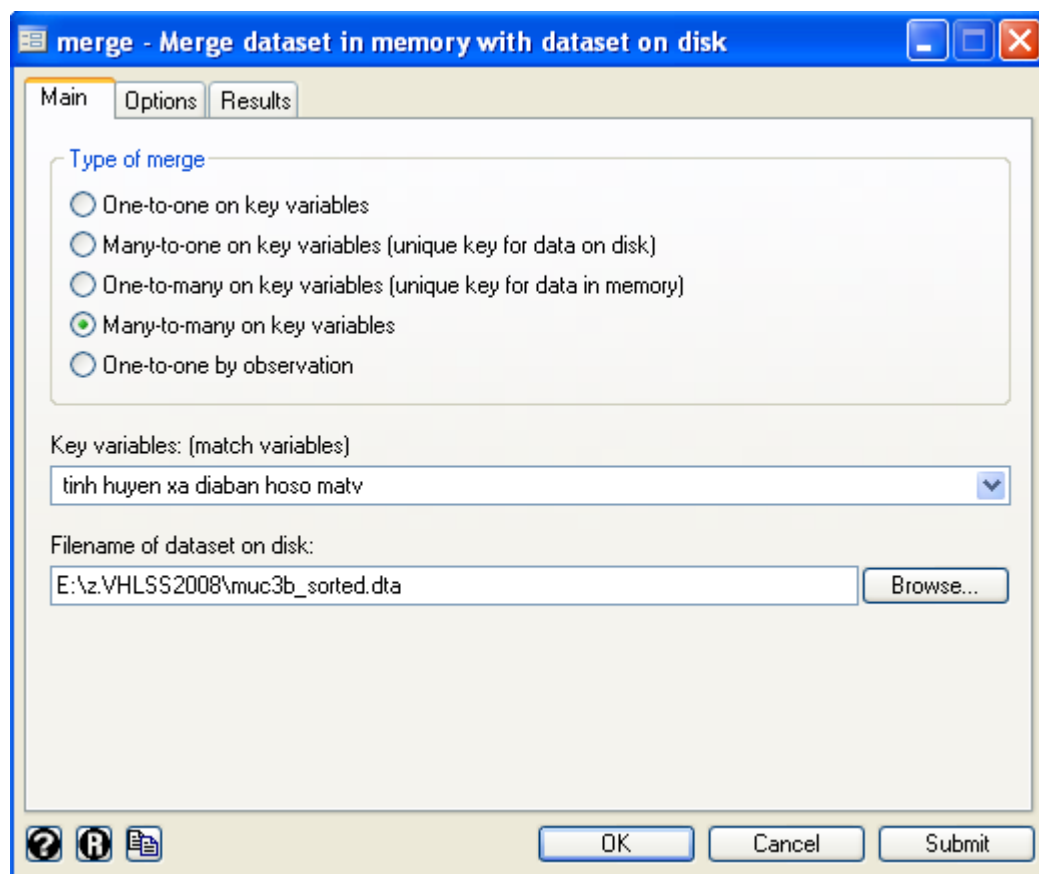
use "E:\MPP5\DATA\VHLSS2008\muc123a.dta", clear

sort tỉnh huyện xa diaban hosomatv

- Thông tin chứa dữ liệu về số lần khám chữa bệnh nội trú, ngoại trú chứa trong file **muc3b**, vậy ta cần merge dữ liệu vào file master **muc123a_sorted.dta** bằng cách:

- Trong Stata, chọn **Data -> Combine datasets -> Merge two datasets.**

Sau đó trong cửa sổ **Merge – Merge Dataset in memory with dataset on disk**, trong thẻ **Main** chọn **Many to many on key variables**; trong mục **Filename of dataset on disk** chọn tới file **muc3b_sorted.dta**



- Ta được kết quả như sau:

merge m:m tinh huyen xa diaban hoso matv using "E:\z.VHLSS2008\muc3b_sorted.dta"

Result	# of obs.	
not matched	27,190	
from master	24,852	(_merge==1)
from using	2,338	(_merge==2)
matched	16,959	(_merge==3)

. keep if *_merge*==3
(27190 observations deleted)

- Để biết dữ liệu về số lượng Nam, Nữ có đi khám chữa bệnh ngoại trú, ta sử dụng lệnh **. tab m3c10a m1ac2 if m3c10a>0**, ta được kết quả như sau:

10. số lần KCB ngoài trú	2. Giới tính		Total
	Nam	Nữ	
1	2,776	3,644	6,420
2	1,415	1,944	3,359
3	787	1,015	1,802
4	460	634	1,094
5	245	295	540
6	188	258	446
7	58	83	141
8	66	71	137
9	15	16	31
10	112	154	266
11	1	8	9
12	104	168	272
13	7	7	14
14	7	14	21
15	21	22	43
16	3	8	11
17	2	1	3
18	3	2	5
19	0	2	2
20	19	37	56
21	2	0	2
22	0	1	1
23	2	2	4
24	19	26	45
25	1	5	6
26	0	1	1
28	0	1	1
30	8	15	23
32	0	2	2
34	1	1	2
35	1	2	3
36	5	9	14
38	0	1	1
40	1	2	3
42	0	2	2
44	0	1	1
45	0	3	3
48	2	6	8
50	0	1	1
52	0	2	2
58	0	1	1
60	0	1	1
80	0	4	4
Total	6,331	8,472	14,803

- Để biết dữ liệu về số lượng Nam, Nữ có đi khám chữa bệnh nội trú, ta sử dụng lệnh `tab m3c12a m1ac2 if m3c12a>0`, ta được kết quả như sau:

. tab m3c12a m1ac2 if m3c12a>0

12. sè lçn nhËp viËn	2. Giíi tÝnh		Total
	Nam	N÷	
1	1,020	1,347	2,367
2	161	131	292
3	36	52	88
4	19	23	42
5	3	5	8
6	2	3	5
7	1	1	2
9	1	1	2
10	2	2	4
12	0	3	3
Total	1,245	1,568	2,813

- Để biết dữ liệu về số lượng Nam vừa đi khám chữa bệnh ngoại trú, vừa đi khám chữa bệnh ngoại trú, ta sử dụng lệnh **tab m3c10a m3c12a if (m3c10a>0) & (m3c12a>0) & m1ac2==1**, ta được kết quả như sau:

. tab m3c10a m3c12a if (m3c10a>0) & (m3c12a>0) & m1ac2==1

10. sè lçn KCB ngo'i trú	12. sè lçn nhËp viËn					Total
	1	2	3	4	7	
1	122	6	0	0	1	129
2	44	25	2	0	0	71
3	19	5	6	1	0	31
4	20	6	0	1	0	27
5	8	1	0	0	0	9
6	5	1	0	1	0	7
7	2	2	0	1	0	5
8	3	0	0	0	0	3
9	0	0	0	1	0	1
10	3	0	1	0	0	4
12	5	1	0	0	0	6
18	1	0	0	0	0	1
20	1	0	0	0	0	1
Total	233	47	9	5	1	295

- Để biết dữ liệu về số lượng Nữ vừa đi khám chữa bệnh ngoại trú, vừa đi khám chữa bệnh ngoại trú, ta sử dụng lệnh **tab m3c10a m3c12a if (m3c10a>0) & (m3c12a>0) & m1ac2==2**, ta được kết quả như sau:

. tab m3<10a m3<12a if (m3<10a>0) & (m3<12a>0) & m1ac2==2

10. sè lçn KCB ngo'í tró	12. sè lçn nhÿp viÿn						Total
	1	2	3	4	5	6	
1	154	9	5	0	0	0	168
2	58	16	1	2	1	0	78
3	30	2	3	1	0	0	36
4	16	4	1	3	0	1	25
5	15	3	4	1	0	0	23
6	7	1	3	0	0	0	11
7	1	0	1	0	0	0	2
8	2	2	0	0	1	0	5
9	1	0	0	0	0	0	1
10	4	1	0	0	0	0	5
11	0	0	0	0	1	0	1
12	4	1	0	0	0	0	5
13	0	0	0	1	0	0	1
16	1	0	0	0	0	0	1
19	1	0	0	0	0	0	1
20	3	1	0	0	0	0	4
24	1	0	0	0	0	0	1
Total	298	40	18	8	3	1	368

-Như vậy ta có thể thống kê được bảng số liệu như sau:

	Nam	% Nam/Tổng số	Nữ	% Nữ/Tổng số	Tổng cộng
1. Ngoại trú	6331	42.77%	8472	57.23%	14803
2. Nội trú	1245	44.26%	1568	55.74%	2813
3. Vừa khám nội trú, vừa khám ngoại trú	295	44.49%	368	55.51%	663
4. Tổng số người đi khám chữa bệnh (=1+2-3)	7281	46.43%	9672	61.39%	16953

❖ **Kiểm định về tính chính xác của việc không có khác biệt giới trong việc khám chữa bệnh**

- Ta tiến hành Kiểm định cho Mẫu Lớn đối với một Tỷ lệ Tổng thể như sau:

Gọi: $p_0 = 50\%$ là tỷ lệ nữ đi khám chữa bệnh.

$$\hat{p} = \frac{9672}{16953} = 0,570518492 \text{ là tỷ lệ lao động nữ đi khám chữa bệnh trong mẫu khảo sát.}$$

$n = 16953$ là số lượng mẫu khảo sát

Giả thuyết $H_0: \hat{p} = p_0$ (tương đương với việc tỷ lệ đi khám chữa bệnh của nam và nữ là như nhau)⁴

Giả thuyết thay thế: $H_a: \hat{p} \neq p_0$

Trị thống kê kiểm định:

⁴ Ở đây học viên cũng có thể đặt giả thuyết kiểm định sự khác biệt giữa tỷ lệ đi khám chữa bệnh của nam và nữ một cách trực tiếp ($H_0: p_{nữ} = p_{nam}; H_a: p_{nữ} \neq p_{nam}$) và thực hiện kiểm định theo cách này.

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.570518 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{16953}}} = 18,36353534$$

Vùng bác bỏ:

Ta có $z_{\alpha/2} = z_{0,025} = 1,96$

$z > z_{\alpha/2}$, vậy ta bác bỏ H_0 .

Do đó, với mức ý nghĩa là 5% thì, có đủ cơ sở để cho rằng tỷ lệ đi khám chữa bệnh của nam và nữ là khác nhau.

Chú ý: Có thể kiểm định thông qua Stata bằng cách sử dụng các lệnh như sau và cho kết quả tương tự:

```
. gen NuKb=1 if (m1ac2==2) & (( m3c10a>0) | (m3c12a>0))
. replace NuKb=0 if (m1ac2==1) & (( m3c10a>0) | (m3c12a>0))
. tab NuKb
```

. tab NuKb

NuKb	Freq.	Percent	Cum.
0	7,281	42.95	42.95
1	9,672	57.05	100.00
Total	16,953	100.00	

. prtest NuKb=0.5

```
One-sample test of proportion                               NuKb: Number of obs = 16953
+-----+-----+-----+-----+-----+
Variable      |      Mean   | Std. Err. |          [95% Conf. Interval]
+-----+-----+-----+-----+-----+
NuKb          |   .5705185 |   .0038018 |   .5630672   .5779698
+-----+-----+-----+-----+-----+
p = proportion(NuKb)                                     z = 18.3635
Ho: p = 0.5

      Ha: p < 0.5          Ha: p != 0.5          Ha: p > 0.5
Pr(Z < z) = 1.0000      Pr(|Z| > |z|) = 0.0000      Pr(Z > z) = 0.0000
```

❖ **Kiểm định về tính chính xác của việc không có khác biệt giới trong việc khám chữa bệnh nội trú:**

- Ta tiến hành Kiểm định cho Mẫu Lớn đối với một Tỷ lệ Tổng thể như sau:

Gọi: $p_0 = 50\%$ là tỷ lệ nữ đi khám chữa bệnh nội trú

$$\hat{p} = \frac{1568}{2813} = 0,557412016 \text{ là tỷ lệ lao động nữ khám chữa bệnh nội trú trong mẫu khảo sát}$$

$n = 2813$ là số lượng mẫu khảo sát

Giả thuyết $H_0: \hat{p} = p_0$

Giả thuyết thay thế: $H_a: \hat{p} \neq p_0$

Trị thống kê kiểm định:

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0,557412 - 0,5}{\sqrt{\frac{0,5(1-0,5)}{2813}}} = 6,090005099$$

Vùng bác bỏ:

Ta có $z_{\alpha/2} = z_{0,025} = 1,96$

Do $z > z_{\alpha/2}$ nên bác bỏ H_0 .

Do đó, với mức ý nghĩa là 5% thì, có đủ cơ sở để cho thấy tỷ lệ đi khám chữa bệnh của nội trú của nam và nữ là khác biệt.

Chú ý: Có thể kiểm định thông qua Stata bằng cách sử dụng các lệnh tương tự như đã làm ở trên.

❖ **Kiểm định về tính chính xác của việc không có khác biệt giới trong việc khám chữa bệnh ngoại trú:**

- Ta tiến hành Kiểm định cho Mẫu Lớn đối với một Tỷ lệ Tổng thể như sau:

Gọi: $p_0 = 50\%$ là tỷ lệ nữ đi khám chữa bệnh ngoại trú

$$\hat{p} = \frac{8472}{14803} = 0,572316422 \text{ là tỷ lệ lao động nữ khám chữa bệnh ngoại trú trong mẫu khảo}$$

sát

$n = 14803$ là số lượng mẫu khảo sát

Giả thuyết H_0 : $\hat{p} = p_0$

Giả thuyết thay thế: H_a : $\hat{p} \neq p_0$

Trị thống kê kiểm định:

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0,572316 - 0,5}{\sqrt{\frac{0,5(1-0,5)}{14803}}} = 17,59712819$$

Vùng bác bỏ:

Ta có $z_{\alpha/2} = z_{0,025} = 1,96$

Do $z > z_{\alpha/2}$ nên bác bỏ H_0 .

Do đó, với mức ý nghĩa là 5% thì, có đủ cơ sở cho thấy tỷ lệ đi khám chữa bệnh ngoại trú của nam và nữ là khác biệt.

Chú ý: Có thể kiểm định thông qua Stata bằng cách sử dụng các lệnh tương tự như đã làm ở trên.

*