Chương Trình Giảng Dạy Kinh tế Fulbright

Học kỳ Thu năm 2011

Các Phương Pháp Phân Tích Định Lượng

Gợi ý giải bài tập 5

KIỂM ĐỊNH THỐNG KÊ

Ngày Phát: Thứ Hai, 31/10/2011
Ngày Nộp: **8:20 sáng, Thứ Hai, 07/11/2011**
Bản in nộp tại Hộp nộp bài tập trong phòng Lab
Bản điện tử gửi lên http://intranet.fetp.edu.vn:81

Bài 1 *(20 điểm)*

Trong năm 1960 một cuộc điều tra dân số đã chỉ ra rằng độ tuổi mà người đàn ông Mỹ lập gia đình lần đầu có trung bình là 23.3. Tuy nhiên, vào năm 201X, có một đánh giá được chấp thuận khá rộng rãi là nam thanh niên ở Mỹ ngày nay lập gia đình trễ hơn. Chúng ta muốn kiểm định xem độ tuổi trung bình của lần lập gia đình đầu tiên của họ có tăng lên sau hơn 50 năm hay không?

a) Hãy lập ra các giả thuyết hợp lý (H_0 và H_a)?

   H_0: $\mu = 23.3$ (độ tuổi trung bình của lần đầu tiên lập gia đình của đàn ông Mỹ là 23.3 tuổi)

   H_a: $\mu > 23.3$ (độ tuổi trung bình của lần đầu tiên lập gia đình của đàn ông Mỹ lớn hơn 23.3 tuổi)

b) Lấy mẫu ngẫu nhiên 120 đàn ông ở Mỹ mới lập gia đình lần đầu tiên năm ngoái. Kết quả thống kê mô tả từ mẫu cho thấy độ tuổi trung bình của họ khi lập gia đình lần đầu tiên là 26.2 tuổi và độ lệch chuẩn là 4.5 năm.Thực hiện việc kiểm định giả thuyết và tìm p_{value}?

   Trị thống kê kiểm định cho cỡ mẫu lớn

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{26.2 - 23.3}{4.5 / \sqrt{120}}$$

   $z = 2.9/0.41 = 7.06$

   Khi mức ý nghĩa là 0.05, giá trị tới hạn 1 phía $z_{0.05}=1.645$

   $p_{value} = P(z > 7.06) = 8.35302E\text{-}13 \simeq 0.000$

Do $z = 7.06 > z_{0.05} = 1.645$ (hay $P_{value} < 0.05$) nên bác bỏ H_0, chấp nhận H_a ở độ tin cậy 95%.

c) Giải thích ý nghĩa của p_{value} trong trường hợp này?

Xác suất mắc sai lầm loại I khi bác bỏ H_0 là 0.000 ; Do xác suất này nhỏ hơn xác suất tối đa mà chúng ta đặt ra (0.05) nên có thể bác bỏ giả thuyết H_0.

d) Các anh chị hãy cho biết kết luận của mình?

Kết luận đưa ra là: tuổi trung bình của lần lập gia đình đầu tiên của người dân hiện nay (năm 201X) đã tăng lên so với năm 1960.

Bài 2 *(20 điểm)*

Một sinh viên điều tra những ảnh hưởng tiềm tàng của chất caffeine đối với việc học thi. Có 40 sinh viên tình nguyện tham gia quá trình điều tra này và được chia làm 2 nhóm. Đầu tiên, mỗi nhóm sẽ làm một bài kiểm tra về trí nhớ. Sau đó, một nhóm sẽ được cho uống hai ly nước ngọt có caffeine và nhóm còn lại sẽ uống hai ly không có caffeine. Ba mươi phút sau mỗi nhóm sẽ thi một bài kiểm tra khác có độ khó tương đương với bài kiểm tra đầu tiên. So sánh kết quả thi trong 2 lần cho thấy đối với nhóm 20 sinh viên uống nước ngọt có chất caffeine thì điểm trung bình giảm xuống 0.925 điểm với độ lệch chuẩn là 2.955 điểm. Đối với nhóm uống nước ngọt không có caffeine thì điểm trung bình tăng lên 1.552 điểm với độ lệch chuẩn là 2.441 điểm. Giả sử khác biệt về điểm thi tuân theo phân phối chuẩn.

a) Điểm của nhóm sinh viên uống nước ngọt có chất caffeine có thay đổi một cách có ý nghĩa không? Kiểm định các giả thuyết và cho biết kết luận của các bạn?

Đây là bài toán kiểm định sự khác biệt cặp giữa hai trung bình tổng thể

H_0: $\mu_d = 0$ (không có sự khác biệt về điểm trung bình của nhóm sinh viên khi uống nước ngọt có chất caffein so với khi không uống)

H_a: $\mu_d \neq 0$ (có sự khác biệt về điểm trung bình của nhóm sinh viên trước và sau khi uống nước ngọt có chất caffein)

Trị thống kê kiểm định cho cỡ mẫu nhỏ:

$$t = \frac{\bar{d} - 0}{s_d \sqrt{n}} = \frac{-0.925}{2.955 / \sqrt{20}} = -1.399$$

$|t| = 1.399 < 2.093$ ➔Không bác bỏ H0

b) Điểm của nhóm sinh viên uống nước ngọt không có chất caffeine có thay đổi một cách có ý nghĩa không? Kiểm định các giả thuyết và cho biết kết luận của các bạn?

Giả thuyết thống kê:

H_0: $\mu_d = 0$ (không có sự khác biệt về điểm trung bình của nhóm sinh viên trước và sau khi uống nước ngọt không có chất caffein)

H_a: $\mu_d \neq 0$ (có sự khác biệt về điểm trung bình của nhóm sinh viên trước và sau khi uống nước ngọt không có chất caffein)

Trị thống kê kiểm định cho cỡ mẫu nhỏ:

$$t = \frac{\bar{d} - 0}{s_d \sqrt{n}} = \frac{1.552}{2.441/\sqrt{20}} = 2.843$$

|t|=2.843 >2.093 ➔bác bỏ Ho, Chấp nhận H_a. Nói cách khác, điểm trung bình của nhóm sinh viên uống nước ngọt không có chất caffeine thay đổi đáng kể so với lúc trước khi uống nước ngọt.

Bài 3 *(20 điểm)*

Một Trung tâm Quốc gia về Thống kê Giáo dục thường xuyên giám sát các khía cạnh khác nhau của giáo dục phổ thông. Các số liệu thống kê của họ trong năm 2005 và 2010 được xem là cơ sở để đánh giá những thay đổi trong giáo dục phổ thông. Trong năm 2005 có 34% học sinh đã không nghỉ học một ngày nào. Một cuộc điều tra trong năm 2010 cho thấy khi khảo sát 9000 học sinh phổ thông thì tỷ lệ này giảm xuống còn 30%. Các nhà quản lý giáo dục rất quan tâm đến vấn đề này và cho rằng tỷ lệ học sinh tham gia lớp đã có sự thay đổi.

a) Hãy lập ra các giả thuyết hợp lý (H_0 và H_a)?

H_0: $p_{2010} = 0.34$ (tỉ lệ học sinh tham gia lớp học không có sự thay đổi)

H_a: $p_{2010} \neq 0.34$ (tỉ lệ học sinh tham gia lớp học có sự thay đổi)

b) Thực hiện việc kiểm định giả thuyết và tìm giá trị p_{value}?

Trị thống kê kiểm định cho cỡ mẫu lớn:

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.3 - 0.34}{\sqrt{\frac{0.34 * 0.66}{9000}}}$$

➔*z=-8.01*

p_{value} = 2*P(z > 8.01) ≃ 0.000

Do |z|=8.01 > 1.96 (hay P-value<0.05) nên bác bỏ H_0, chấp nhận H_a ở độ tin cậy 95%.

c) Các bạn có kết luận gì về tình trạng tham gia lớp của các học sinh?

Tình trạng tham gia lớp của học sinh đã thay đổi một cách đáng kể so với năm 2005.

d) Bạn có nghĩ rằng sự khác biệt giữa tỷ lệ học sinh tham gia lớp trong hai năm 2010 và 2005 là có ý nghĩa không? Giải thích?

Sự khác biệt giữa tỷ lệ học sinh tham gia lớp trong hai măm 2010 và 2005 là có ý nghĩa về mặt thống kê. Nếu như không có vấn đề tiêu cực, hay "bệnh thành tích trong giáo dục" thì việc giảm tỷ lệ này là đáng lo ngại cho những người quản lý giáo dục.

Bài 4 *(20 điểm)*.

Kết quả điều tra mức sống hộ gia đình Việt Nam năm 2008 có thông tin việc làm của trẻ em từ 6 đến dưới 15 tuổi. Giả sử mẫu được chọn theo phương pháp ngẫu nhiên. Trong số 2991 em nam có 251 em đang làm việc; và có 222 em nữ đang làm việc trong 2907 em nữ được điều tra. Liệu có

đủ bằng chứng để kết luận rằng tỷ lệ trẻ em nam đang làm việc khác với tỷ lệ trẻ em nữ đang làm việc hay không?

Tỷ lệ lao động ở trẻ em:

$p^{\wedge}_{nam} = 251/2991 = 0.084$; $q^{\wedge}_{nam} = 1 - 0.084 = 0.916$

$p^{\wedge}_{nữ} = 222/2907 = 0.076$; $q^{\wedge}_{nữ} = 1 - 0.076 = 0.924$

Giả thuyết thống kê:

H_0: $p_{nam} - p_{nữ} = 0$ (tỉ lệ trẻ em nam và nữ đang làm việc không khác nhau)

H_a: $p_{nam} - p_{nữ} \neq 0$ (tỉ lệ trẻ em nam đang làm việc khác với tỉ lệ trẻ em nữ đang làm việc)

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}} = \frac{(\hat{p}_{nam} - \hat{p}_{nu}) - D_0}{\sqrt{\frac{\hat{p}_{nam}\hat{q}_{nam}}{n_1} + \frac{\hat{p}_{nu}\hat{q}_{nu}}{n_2}}} = \frac{0.084 - 0.076}{\sqrt{\frac{0.084*0.916}{2991} + \frac{0.076*0.924}{2907}}}$$

$z = 1.068 \rightarrow |z| < 1.96$ ➔ Không bác bỏ giả thuyết H_0 với mức ý nghĩa 5%

➔ Chưa đủ cơ sở để bác bỏ giả thuyết H_0, hay sự khác biệt về tỷ lệ lao động trẻ em giữa nam và nữ là không đáng kể (không có ý nghĩa thống kê).

Bài 5 *(20 điểm)* Bạn hãy sử dụng phần mềm Stata và dữ liệu VHLSS2008, tính toán các kết quả cần thiết để trả lời các câu hỏi sau (để đơn giản, bạn chưa cần tính đến yếu tố trọng số). Dưới đây là một số gợi ý, kết quả của bạn có thể khác một chút, tùy mục tiêu, giới hạn, giả định của bạn.

a. Có người cho rằng: chi tiêu trung bình cho việc học tập trong một năm của học sinh bậc trung học phổ thông là từ 2 triệu đồng trở lên (tạm thời sử dụng biến m2ac13k để đo lường chi tiêu cho việc học tập; chỉ xét những người đang đi học hoặc đang nghỉ hè tại thời điểm điều tra). Theo bạn, phát biểu này là đúng hay sai?

```
. ttest m2ac13k=2000 if (m2ac5<=2) & ( m2ac8==3)

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 m2ac13k |    1693    1633.879    43.97914    1809.569     1547.62    1720.138
------------------------------------------------------------------------------
    mean = mean(m2ac13k)                                          t =  -8.3249
Ho: mean = 2000                                 degrees of freedom =     1692

   Ha: mean < 2000             Ha: mean != 2000                 Ha: mean > 2000
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Ta thấy, kiểm định giả thuyết 1 phía H_0: Trung bình>=2000 bị bác bỏ do P-value <0.05. Phát biểu của tác giả trên là không đúng.

b. Có sự khác biệt về chi tiêu trung bình cho việc học tập của học sinh trung học phổ thông giữa trường công lập và ngoài công lập (bao gồm bán công, dân lập, tư thục, và hệ khác) hay không? Nếu có, hãy ước lượng sự khác biệt này?

```
. gen conglap=m2ac9==1
```

```
. tab conglap

    conglap |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     26,194       74.51       74.51
          1 |      8,960       25.49      100.00
------------+-----------------------------------
      Total |     35,154      100.00

. tab conglap if (m2ac5<=2) & ( m2ac8==3)

    conglap |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        272       16.07       16.07
          1 |      1,421       83.93      100.00
------------+-----------------------------------
      Total |      1,693      100.00

. display 188+66+14+4
272


. robvar m2ac13k if (m2ac5<=2) & ( m2ac8==3), by(conglap)

            | Summary of 13k.Tæng sè (a+b+...+i)
    conglap |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   2578.1691   3019.0531         272
          1 |   1453.1281   1399.5677        1421
------------+------------------------------------
      Total |   1633.8789   1809.5693        1693

W0  =  42.793877   df(1, 1691)     Pr > F = 0.00000000

W50 =  25.466884   df(1, 1691)     Pr > F = 0.00000050

W10 =  26.857094   df(1, 1691)     Pr > F = 0.00000025
```

Do P-value của thống kê W0 trong kiểm định về sự bằng nhau của phương sai =0.000. Nên phương sai giữa 2 nhóm là khác nhau.

```
. ttest m2ac13k if (m2ac5<=2) & ( m2ac8==3), by(conglap) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     272    2578.169     183.057    3019.053    2217.775    2938.564
       1 |    1421    1453.128     37.1276    1399.568    1380.297    1525.959
---------+--------------------------------------------------------------------
combined |    1693    1633.879    43.97914    1809.569     1547.62    1720.138
---------+--------------------------------------------------------------------
    diff |            1125.041    186.7841                757.4358    1492.646
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   6.0232
Ho: diff = 0                     Satterthwaite's degrees of freedom =  293.659

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Chi tiêu cho giáo dục ở bậc học trung học phổ thông có sự khác biệt giữa những trường công lập và ngoài công lập. Chi tiêu của nhóm ngoài công lập cao hơn so với công lập. Ở độ tin cậy 95%, sự khác biệt này là từ 757.4 ngàn đồng đến 1492.6 ngàn đồng.

c. Có sự khác biệt giữa trung bình chi cho việc học thêm các môn học thuộc chương trình (biến m2ac13h) và trung bình chi học phí (biến m2ac13a) của học sinh trung học phổ thông ở khu vực thành thị hay không? Sự khác biệt này như thế nào?

```
. ttest m2ac13h =m2ac13a if m2ac5<=2 &  m2ac8==3 & urban08==1 & m2ac13h!=-2 &
m2ac13a!=-2
```

```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 m2ac13h |     428    1019.498    86.18192    1782.945    850.1041    1188.891
 m2ac13a |     428    563.7757    62.00313    1282.731    441.9064     685.645
---------+--------------------------------------------------------------------
    diff |     428     455.722    106.3369    2199.914    246.7131    664.7308
------------------------------------------------------------------------------
     mean(diff) = mean(m2ac13h - m2ac13a)                          t =   4.2856
 Ho: mean(diff) = 0                           degrees of freedom =      427

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Có sự khác biệt trung bình chi cho việc học thêm các môn học thuộc chương trình (biến m2ac13h) và trung bình chi học phí (biến m2ac13a) của học sinh trung học phổ thông ở khu vực thành thị hay không (Do chấp nhận giả thuyết Ha: mean(diff) > 0). Và chi tiêu cho việc học thêm các môn học thuộc chương trình cao hơn so với chi học phí.

d. Có sự khác biệt về tỷ lệ người có việc làm (chỉ xét những người từ 15 tuổi trở lên ở khu vực thành thị) giữa Vùng Đông Nam Bộ và Vùng Đồng Bằng Sông Hồng hay không (nếu xét ở mức ý nghĩa 10%)?

```
. gen covieclam= m4ac2
. recode covieclam 2=0
(covieclam: 12574 changes made)

. tab covieclam

  covieclam |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     12,574       35.77       35.77
          1 |     22,580       64.23      100.00
------------+-----------------------------------
      Total |     35,154      100.00

. tab covieclam if (m1ac5>=15) & (urban08==1) & (reg8==1 | reg8==7)

  covieclam |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,150       36.12       36.12
          1 |      2,034       63.88      100.00
------------+-----------------------------------
      Total |      3,184      100.00

. prtest covieclam if (m1ac5>=15) & (urban08==1) & (reg8==1 | reg8==7), by (reg8)

Two-sample test of proportion                     1: Number of obs =     1324
                                                  7: Number of obs =     1860
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           1 |   .6200906    .013339                      .5939467    .6462346
           7 |   .6521505   .0110437                      .6305054    .6737957
-------------+----------------------------------------------------------------
        diff |  -.0320599   .0173174                     -.0660013    .0018815
             |  under Ho:   .0172718    -1.86   0.063
------------------------------------------------------------------------------
        diff = prop(1) - prop(7)                                  z =  -1.8562
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0317         Pr(|Z| < |z|) = 0.0634          Pr(Z > z) = 0.9683
```

Do P-value của kiểm định 2 phía bằng 0.0634 (<0.1), nên ở độ tin cậy 90%, có sự khác biệt về tỷ lệ có việc làm ở thành thị giữa Vùng ĐB Sông Hồng và Vùng Đông Nam Bộ.

e* (điểm cộng). Bạn hãy thực hiện lại các kiểm định ở câu b, câu c nếu như các giả định của chúng không được thỏa mãn. Bạn hãy tiến hành kiểm định các giả định có được thỏa mãn hay

không. Nếu các giả định bị vi phạm bạn hãy sử dụng một kiểm định phi tham số phù hợp để thay thế.

Ở câu b:

NÕu c¸c gi¶ ®Þnh cña KiÓm ®Þnh T cho 2 mÉu ®éc lËp kh«ng tháa m·n, b¹n h·y sö dông kiÓm ®Þnh Mann-Whitney U - còng cßn gäi lµ kiÓm ®Þnh tæng-h¹ng Wilcoxon(Hamilton,2006).Nh÷ng tr-êng hîp sau ®©y b¹n cã thÓ ¸p dông kiÓm ®Þnh Wilcoxon: biÕn cÇn kiÓm ®Þnh cã thang ®o thø bËc, gi¶ ®Þnh ph©n phèi chuÈn kh«ng tháa m·n, cì mÉu nhá, cã nh÷ng vÊn ®Ò do c¸c gi¸ trÞ bÊt th-êng g©y nªn

Ở câu c:

KiÓm ®Þnh phi tham sè thay thÕ: signrank (kiÓm ®Þnh dÊu vµ h¹ng)