Chương Trình Giảng Dạy Kinh tế Fulbright

Học kỳ Thu năm 2014

Các Phương Pháp Phân Tích Định Lượng

ĐÁP ÁN BÀI TẬP 5

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

Ngày Phát: **Thứ ba 28/10/2014**
Ngày Nộp: **Thứ ba 04/11/2014**
Bản in nộp lúc 8h20 tại Hộp nộp bài tập trong phòng Lab
Bản điện tử gửi lên http://www.fetp.edu.vn/vn/tai-nguyen/hoc-vien-hien-tai/

Câu 1: (40đ)

Vào tháng 10/2012, anh T đã xem xét thông tin về tuổi của học viên MPP khi vừa nhập học (từ MPP1 đến MPP5). Sau khi tính toán, anh T có được 02 kết quả như sau:

Kết quả 1: tuổi trung bình của học viên MPP khi vừa nhập học là 26 tuổi.

Kết quả 2: tuổi trung bình của học viên nam MPP khi vừa nhập học lớn hơn 27 tuổi.

Hiện nay, danh sách học viên MPP đã bổ sung thêm 2 khóa là MPP6 và MPP7, anh T không biết 02 kết luận trên của anh vào năm 2012 có còn đúng tại thời điểm này hay không? Do vậy, anh T đã tiến hành lấy mẫu ngẫu nhiên về tuổi của 54 học viên MPP khi vừa nhập học trong tổng số khoảng 500 học viên MPP các khoá từ MPP1 đến MPP7 để kiểm định lại 02 kết quả trên.

*Thông tin về tuổi của 54 học viên trong file **age.xls** đính kèm.*

Dựa vào các thông tin và dữ liệu được cung cấp, anh/chị hãy trả lời và đưa ra các nhận định từ kết quả tính toán cho các câu hỏi sau đây:

a. Lập ra các giả thuyết hợp lý để giúp anh T kiểm định Kết quả 1:

$H_0 : \mu = 26$ (tuổi trung bình của học viên MPP khi vừa nhập học FETP là 26 tuổi)

$H_a : \mu \neq 26$ (tuổi trung bình của học viên MPP khi vừa nhập học không phải là 26 tuổi)

b. Thực hiện việc kiểm định giả thuyết trên với mức ý nghĩa α = 5% và tìm giá trị p_{value}?

Thực hiện TK mô tả cho n = 54 học viên, $\bar{x} = 27.72$ **tuổi** và $\sigma = 3.43$ tuổi

Vì n = 54 > 30 => cỡ mẫu lớn, có thể áp dụng trị thống kê Z để kiểm định

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{27.72 - 26}{3.43/\sqrt{54}} = 3.69$$

Với α = 5%, vì là kiểm định 2 phía, nên $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$

(Có thể dùng excel hoặc tra bảng Z để tìm $Z_{\alpha/2} = 1.96$)

Vậy $Z = 3.69 > Z_{\alpha/2} = 1.96$ => bác bỏ giả thuyết H_0 và chấp nhận giả thuyết H_a.

Với độ tin cậy 95%, có thể kết luận rằng tuổi trung bình của học viên MPP khi vừa nhập học khác 26 tuổi. Hay tuổi trung bình của học viên MPP khi vừa nhập học khác 26 tuổi một cách có ý nghĩa thống kê ở mức α = 5%.

Như vậy, tuổi trung bình của học viên MPP bây giờ không phải là 26 tuổi.

- **Tìm p_{value}:** là mức ý nghĩa quan sát được của kiểm định, đó là giá trị nhỏ nhất của α mà qua đó các kết quả kiểm định là có ý nghĩa về mặt thống kê.

  $p_{value} = P(Z > 3.69) + P(Z < -3.69) = 0.02\% < 5\%$ => báo bỏ giả thuyết H_0, chấp nhận H_a.

Như vậy kết quả kiểm định bằng thống kê Z hay p_{value} đều cho kết quả tương tự nhau.

c. Tính khoảng tin cậy cho tuổi trung bình của học viên MPP với mức ý nghĩa α = 5%?

$$\left.\begin{aligned} \bar{x} - Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = 3.69 - 1.96 * \frac{3.43}{\sqrt{54}} = 26.81 \\ \bar{x} + Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = 3.69 + 1.96 * \frac{3.43}{\sqrt{54}} = 28.64 \end{aligned}\right\}$$ => Khoảng tin cậy: [26.81 ; 28.64] tuổi

Vì $\mu_0 = 26 \notin [26.81; 28.64]$ tuổi => bác bỏ giả thuyết H_0, chấp nhận giả thuyết H_a ở mức ý nghĩa **α = 5%.** Kết quả cũng tương tự như câu a và b.

d. Từ kết quả tính toán ở trên, nếu anh T vẫn cho rằng kết quả 1 mà mình tính toán năm 2012 vẫn đúng cho hiện tại thì anh T đang mắc sai lầm loại nào?

Kết quả từ các câu trên cho thấy khi có thêm học viên từ lớp MPP6 và MPP7, tổng thể học viên MPP đã thay đổi và tuổi trung bình của học viên MPP bây giờ khác 26. Nếu kết luận rằng tuổi trung bình của MPP vẫn là 26, tức là anh T đang chấp nhận cái sai. Như vậy anh T mắc sai lầm loại 2.

e. Lập các giả thuyết hợp lý để giúp anh T kiểm định Kết quả 2?

$H_0 : \mu \geq 27$ (tuổi TB của học viên nam MPP khi vừa nhập học lớn hơn hoặc bằng 27 tuổi)

$H_a : \mu < 27$ (tuổi TB của học viên MPP khi vừa nhập học nhỏ hơn 27 tuổi)

f. Thực hiện việc kiểm định giả thuyết trên

Từ dữ liệu thu thập, đếm số quan sát là tuổi của học viên nam (dùng hàm countif với gender =1) và các hàm thống kê mô tả, kết quả cho **n = 28 học viên,** $\bar{x} = 27.43$ **tuổi** và $\sigma = 2.95$ tuổi.

Kết quả từ lấy mẫu ngẫu nhiên cho thấy, trung bình tuổi học viên nam trong mẫu thu thập là 27.43 tuổi, lớn hơn 27 tuổi. Chúng ta sẽ không thực hiện kiểm định vì kết quả kiểm định sẽ luôn luôn là không thể bác bỏ H_0. Mặt khác, khi kết quả từ mẫu thu thập không khác với phát biểu về tổng thể thì chúng ta không còn lý do gì để nghi ngờ và đi kiểm định tổng thể.

KẾT LUẬN:

Như vậy, với độ tin cậy 95% cho thấy: (i) tuổi học viên MPP khi vừa nhập học không còn là 26 tuổi và (ii) chúng ta chưa có đủ cơ sở để cho rằng tuổi trung bình của học viên nam MPP khi vừa nhập học không còn lớn hơn hoặc bằng 27 tuổi.

Câu 2: (20đ)

Sử dụng file Employee_Data.xls để làm bài.

Trong Bài tập 1, mặc dù không thể tính hệ số tương quan giữa Biến **salary** và Biến **gender**, tuy nhiên, một số nhận định đã cho rằng, lương trung bình của nam sẽ cao hơn nữ bởi nam thì có sức khỏe hơn, có thể làm nhiều việc hơn nên lương cao hơn?!

Dựa vào kết quả thu thập một cách ngẫu nhiên thông tin về salary và gender của 474 lao động trong ngành viễn thông anh chị hãy trả lời các câu hỏi sau đây:

a. Lập giả thuyết hợp lý cho nhận định trên?

$H_0 : \mu_1 \geq \mu_2$ *hay* $\mu_1 - \mu_2 \geq D_0 = 0$ (lương TB của nam cao hơn so với lương TB của nữ)

$H_a : \mu_1 < \mu_2$ *hay* $\mu_1 - \mu_2 < D_0 = 0$ (lương TB của nam thấp hơn so với lương TB nữ)

b. Thực hiện việc kiểm định giả thuyết với mức ý nghĩa α = 5% và đưa ra kết luận?

Để có thể kiểm định sự khác biệt về trung bình lương của 2 nam và nữ, dùng hàm countif để tách thành lương của 2 nhóm riêng biệt. Dùng thống kê mô tả ta có được 1 số đặc trưng thống kê cần dùng như sau:

Đặc trưng thống kê	Lương của nam	Lương của nữ
Trung bình ($\bar{x}$) USD/năm	41,442	26,032
Độ lệch chuẩn (σ) USD/năm	19,499	7,558
Số giá trị quan sát (n) giá trị	n_1 = 258	n_2 = 216

Từ các đặc trưng thống kê trên, ta thấy mẫu ngẫu nhiên trong tay ủng hộ H_0: không cần kiểm định, kết quả luôn luôn là không đủ cơ sở bác bỏ H_0.

Vậy nhận định cho rằng lương của nam cao hơn lương của nữ là có cơ sở.

c. Tính p_{value} và kết luận. Với mức ý nghĩa α = 1% thì kết luận ở câu b còn đúng không?

Để có thể tính được p_value, ta phải tính được trị thống kê z hoặc t.

$$Z = \frac{\overline{x_1} - \overline{x_2} - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{41{,}442 - 26{,}032 - 0}{\sqrt{\frac{19{,}499^2}{258} + \frac{7{,}558^2}{216}}} = 11.69$$

$p_{value} = P(Z < 11.69) \approx 1 > 5\%$ => không đủ cơ sở bác bỏ H_0

Như vậy kết quả kiểm định bằng trị thống kê Z hay p_{value} cũng tương tự nhau.

$p_{value} \approx 1$, do vậy, dù ở mức ý nghĩa nào thi cũng không đủ cơ sở bác bỏ H_0.

d. Để phản biện nhận định *"người nam có sức khỏe hơn nên khả năng làm việc nhiều hơn dẫn đến lương cao hơn"*, một số ý kiến cho rằng: đối với công việc làm quản lý, thì sự khác biệt về

lương trung bình của người nam và nữ sẽ không còn. Với giả định phương sai về lương của nam và nữ làm quản lý là như nhau.

Bằng giả thuyết và kiểm định hợp lý, anh/chị hãy đưa ra nhận xét về ý kiến trên trên.

Phát biểu giả thuyết:

$H_0 : \mu_1 = \mu_2$ *hay* $\mu_1 - \mu_2 = D_0 = 0$ (lương trung bình của nam là quản lý không khác biệt so với lương trung bình của nữ là quản lý)

$H_a : \mu_1 \neq \mu_2$ *hay* $\mu_1 - \mu_2 \neq D_0 = 0$ (lương trung bình của nam là quản lý và lương trung bình của nữ là quản lý có sự khác biệt)

Kiểm định giả thuyết

Tương tự, tạo ra 2 nhóm lương của nam là quản lý (gender=1 và jobcat=3) và nhóm lương của nữ là quản lý (gender =2 và jobcat = 3) bằng hàm countif.

Xử lý thống kê mô tả cho kết quả như sau:

Đặc trưng thống kê	Lương của nam	Lương của nữ
Trung bình ($\bar{x}$) USD/năm	66,243	47,214
Độ lệch chuẩn (σ) USD/năm	18,052	8,051
Số giá trị quan sát (n) giá trị	n1 = 74	n2 = 10

Từ bảng trên cho thấy $n_1 = 74 > 30$ nhưng $n_2 = 10 < 30$, do vậy áp dụng thống kê t để kiểm định giả thuyết với giả định phương sai của tổng thể là bằng nhau:

$$t = \frac{\overline{x_1} - \overline{x_2} - D_0}{s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{66{,}243 - 47{,}214 - 0}{17{,}263 * \sqrt{\frac{1}{74} + \frac{1}{10}}} = 3.27$$

Với $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{(74-1) * 18{,}052^2 + (10-1) * 8{,}501^2}{(74 + 10 - 2)}} = 17{,}263$ (USD/năm)

Với **α = 5%,** vì là kiểm định 2 phía, nên $t_{\alpha/2} = t(0.025, n = 74 + 10 - 2) = t(0.02;82) = 1.989$

t = 3.27 > $t_{\alpha/2}$=1.989 => bác bỏ giả thuyết H_0, chấp thuận giả thuyết H_a ở mức **α = 5%,** nghĩa là, trung bình lương của người nam là quản lý vẫn cao hơn lương trung bình của người nữ là quản lý một cách có ý nghĩa thống kê ở mức **α = 5%.**

Câu 3: (20đ)

Thị trường bảo hiểm nhân thọ đã chính thức ra đời vào tháng 08/1996 với sản phẩm đầu tiên của Bảo Việt. Thị trường bảo hiểm sau đó phát triển nhanh chóng trong khoảng 8 năm (1996 – 2004). Tuy nhiên, kể từ năm 2005, thị trường bảo hiểm liên tục gặp khó khăn do nhiều yếu tố tác động (tính mới không còn, khủng hoảng kinh tế, lãi suất ngân hàng tăng,…). Nhưng quan trọng nhất là niềm tin của khách hàng vào bảo hiểm nhân thọ. Một số chuyên gia, lãnh đạo của ngành bảo hiểm cho rằng, tạo dựng niềm tin của khách hàng là điều kiện tiên quyết cho sự tồn tại và phát triển của doanh nghiệp bảo hiểm[1]. Do vậy, nhiều ý kiến cho rằng chính niềm tin của khách hàng sụt giảm là nguyên nhân dẫn đến tình trạng khó khăn của doanh nghiệp bảo hiểm hiện nay. *Dẫn chứng cụ thể, trước đây, tỷ lệ khách hàng đồng ý cho nhân viên bảo hiểm tư vấn sau lần tiếp xúc đầu tiên vào khoảng 20%. Tuy nhiên, hiện nay, tỷ lệ này hiện nay có xu hướng giảm (1 tháng, nhân viên tư vấn phải tiếp xúc và gọi điện cho khoảng 200 khách hàng ngẫu nhiên, nhưng chỉ có khoảng 20 người đồng ý tiếp xúc để được tư vấn).*

Giả định rằng tỷ lệ khách hàng đồng ý tiếp xúc để được tư vấn sau lần gặp đầu tiên càng cao thì niềm tin của khách hàng vào thị trường bảo hiểm càng cao. Với thông tin từ dẫn chứng như trên, anh/chị hãy trả lời các câu hỏi sau đây:

a. Lập ra các giả thuyết hợp lý để kiểm định lại ý kiến trên?

Gọi: p là tỷ lệ của tổng thể, $\hat{p}$ là tỷ lệ của mẫu (= 20/200 = 10%)

p_o là giá trị cụ thể của giả thuyết đối với tỷ lệ của tổng thể.

$p_o = 20\% \Rightarrow q_o = 1 - 20\% = 80\%$

$H_0 : p \geq p_0 = 20\%$ (tỷ lệ khách hàng đồng ý tiếp xúc để được tư vấn sau lần gặp đầu tiên hiện nay không giảm so với trước đây).

$H_a : p < p_0 = 20\%$ (tỷ lệ khách hàng đồng ý tiếp xúc để được tư vấn sau lần gặp đầu tiên hiện nay đã giảm so với trước đây).

b. Thực hiện việc kiểm định giả thuyết với mức ý nghĩa α = 5%?

Vì n = 200 > 30, cỡ mẫu lớn => áp dụng trị thống kê kiểm định như sau:

[1] http://www.libertyinsurance.com.vn/index.php/vi/tin-tuc/tin-tuc/tin-tuc-thi-truong-bao-hiem-viet-nam/215-co-hoi-thi-truong-duoi-goc-nhin-cua-cac-ceo-nganh-bao-hiem?layout=detail

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 * q_0}{n}}} = \frac{20/200 - 20\%}{\sqrt{\frac{20\% * 80\%}{200}}} = -3.54$$

Với **α = 5%,** vì là kiểm định 1 phía, nên $Z_\alpha = Z_{0.05} = 1.645$

(Có thể dùng Excel hoặc tra bảng Z đểm tìm $Z_{0.05}$ = 1.645)

Vậy z = -3.54 < $-Z_\alpha$ = -1.645 => bác bỏ H_0, chấp nhận H_a với mức **α = 5%,** nghĩa là tỷ lệ khách hàng đồng ý tiếp xúc để được tư vấn sau lần gặp đầu tiên giảm so với trước đây một cách có ý nghĩa thống kê ở mức 5%. Hay nói cách khác, niềm tin của khách hàng vào bảo hiểm nhân thọ đã giảm đi có ý nghĩa thống kê với độ tin cậy 95%.

c. Tìm p_{value}? Theo anh/chị, nhận định cho rằng niềm tin của khách hàng vào thị trường bảo hiểm đang giảm sút như trên là có ý nghĩa hay không? Mức độ tin cậy của nhận định trên là bao nhiêu?

Vì là kiểm định một phía bên trái, do vậy p_{value} = P(Z<-3.54) = 0.02% < 1%. Vậy với độ tin cậy (1 - 0.02% = 99.98%), chúng ta có thể kết luận, niềm tin của khách hàng vào bảo hiểm nhân thọ đã giảm đi so với trước đây một các có ý nghĩa thống kê.

<u>Câu 4</u>: (20đ)

Để chuẩn bị cho việc thử nghiệm một chính sách mới về an sinh xã hội, A đã nhận được sự đồng ý tham gia của 445 người có các điều kiện tương tự nhau về tuổi, trình độ học vấn,... Để có thể đo lường được tác động của chính sách mà A chuẩn bị áp dụng, A phải phân chia 445 người trên làm 2 nhóm sao cho tỷ lệ thất nghiệp trong mỗi nhóm là như nhau. Bằng cách phân nhóm ngẫu nhiên, A đã phân được 2 nhóm cụ thể như sau: nhóm 1 bao gồm 185 người sẽ được tham gia chính sách mới, số người còn lại được xếp vào nhóm 2 để làm đối chứng.

File dữ liệu **experiment.xls** ghi nhận thông tin tóm lược về 445 người trên. Trong đó: Biến **thamgia** cho biết người đó được xếp vào nhóm 1 (nếu nhận giá trị là 1) hay được xếp vào nhóm 2 hay nhóm đối chứng (nếu nhận giá trị là 0); Biến **thatnghiep** cho biết người đó đang thất nghiệp (nếu nhận giá trị là 1) và người đó không thất nghiệp (nếu nhận giá trị là 0).

Anh/chị hãy kiểm định xem tỷ lệ người thất nghiệp trong nhóm tham gia (nhóm 1) và nhóm đối chứng (nhóm 2) có khác biệt hay không với mức ý nghĩa α = 5%?

Từ dữ liệu đề bài cho, sử dụng công cụ Pivot Table trong excel, lập bảng tần suất của biến tham gia và biến thất nghiệp, ta có kết quả từ excel như sau:

Count of Thất nghiệp (Y/N)	**Thất nghiệp (1:Có; 0: Không)**		
Tham gia (1:TG; 0: Không TG)	**0**	**1**	**Tổng**
0	82	178	260
1	74	111	185
Tổng	**156**	**289**	**445**

Từ bảng trên ta xác định các trị thống kê cần thiết để kiểm định như sau:

$\hat{p}_2 = 178/260 = 68.46\%$	Tỷ lệ thất nghiệp trong nhóm đối chứng (nhận giá trị là 0)
$\hat{q}_2 = 1 - \hat{p}_2 = 38.41\%$	Tỷ lệ người không thất nghiệp trong nhóm đối chứng.
$n_2 = 260$ người	Tổng số người nhóm đối chứng.
$\hat{p}_1 = 111/185 = 60.00\%$	Tỷ lệ thất nghiệp trong nhóm tham gia (nhận giá trị là 1)
$\hat{q}_1 = 1 - \hat{p}_1 = 40.00\%$	Tỷ lệ người không thất nghiệp trong nhóm tham gia.
$n_1 = 185$ người	Tổng số người nhóm tham gia.

Phát biểu giả thuyết:

H_0: $p_1 - p_2 = 0$ (tỷ lệ thất nghiệm giữa nhóm đối chứng và nhóm tham gia không có sự khác biệt).

H_a: $p_1 - p_2 \neq 0$ (tỷ lệ thất nghiệm giữa nhóm đối chứng và nhóm tham gia có sự khác biệt).

Vì $n_2 = 260 > 30$ và $n_1 = 185 > 30$, cỡ mẫu lớn, dùng trị thống kê để kiểm định sự khác biệt giữa hai tỷ lệ nhị thức như sau:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}} = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\frac{\hat{p}_1 * \hat{q}_1}{n_1} + \frac{\hat{p}_2 * \hat{q}_2}{n_2}}} = \frac{60.00\% - 68.46\% - 0}{\sqrt{\frac{60.00\% * 40\%}{185} + \frac{68.46\% * 31.54\%}{260}}} = -1.45$$

Với **α = 5%,** vì là kiểm định 2 phía, nên $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$

Vì $Z = -1.45 \in [-Z_{\alpha/2}; Z_{\alpha/2}] = [-1.96; 1.96]$ => chưa đủ cơ sở bác bỏ H_0, ở mức **α = 5%,** nghĩa là với độ tin cậy 95, chúng ta chưa đủ cơ sở để nói rằng tỷ lệ thất nghiệp giữa nhóm đối chứng và nhóm tham gia có sự khác bmột cách có ý nghĩa thống kê.