Chương trình Giảng dạy Kinh tế Fulbright

Học kỳ Thu năm 2014

Các Phương Pháp Phân Tích Định Lượng

GỢI Ý ĐÁP ÁN BÀI TẬP 4

XÁC SUẤT
Ngày Phát: Thứ Ba 21/10/2014
Ngày Nộp: Thứ Ba 28/10/2014

Bài 1: (20 điểm)

Một cửa hàng thương mại ghi nhận doanh thu trong 5 tháng như sau: 300; 350; 250; 300; 400 (triệu đồng). Giả sử doanh thu trong 5 tháng trên đây được xem như một tổng thể. Tiến hành chọn mẫu ngẫu nhiên 2 trong 5 tháng và tìm hiểu về doanh thu trung bình của cửa hàng trong 2 tháng.

a. Tính tham số thống kê: trung bình, độ lệch chuẩn của tổng thể.

b. Tìm phân phối xác suất của trung bình mẫu $\overline{X}$.

c. Tính và nêu ý nghĩa các trị thống kê: trung bình, độ lệch chuẩn trực tiếp từ phân phối xác suất của $\overline{X}$.

d. Chỉ ra và giải thích các mối quan hệ giữa trị thống kê và tham số thống kê nói trên.

Đáp án bài 1:

Gọi doanh thu của 5 tháng lần lượt là: x_1, x_2, x_3, x_4, x_5

Chọn ngẫu nhiên doanh thu của 2 trong 5 tháng

a. Tính tham số thống kê: trung bình, độ lệch chuẩn của tổng thể.

Xem doanh thu trong 5 tháng là một tổng thể:

$\mu_X = \sum x_i/N = (x_1 + x_2 + x_3 + x_4 + x_5)/5 = (300 + 350 + 250 + 300 + 400)/5 = 320$ (triệu đồng)

$\sigma_X^2 = \sum(x_i - \mu)^2/N = \frac{(300-320)^2+(350-320)^2+(250-320)^2+(300-320)^2+(400-320)^2}{5} = 2600$ (triệu đồng)^2

$\sigma_X = \sqrt{\sigma_X^2} = 50{,}99$ (triệu đồng)

b. Tìm phân phối xác suất của trung bình mẫu $\overline{X}$.

Với tổng thể 5 phần tử trên, ta có tất cả C_5^3=10 cách chọn mẫu gồm 2 phần tử và tính ra được trung bình mẫu $\overline{X}$ theo bảng như sau:

Mẫu thứ	1	2	3	4	5	6	7	8	9	10
x_1	300				300	300	300			
x_2	350	350						350	350	
x_3		250	250		250					250
x_4			300	300		300		300		

x_5				400			400		400	400
$\overline{X}$	325	300	275	350	275	300	350	325	375	325

Vậy phân phối xác suất trung bình của mẫu $\overline{X}$ là:

$\overline{X}$	275	300	325	350	375
Tần suất P($\overline{X}$)	2/10	2/10	3/10	2/10	1/10

c. Tính và nêu ý nghĩa các trị thống kê: trung bình, độ lệch chuẩn trực tiếp từ phân phối xác suất của $\overline{X}$.

Áp dụng công thức tính trung bình mẫu:

$\mu_{\overline{X}} = \sum \overline{X}.P(\overline{X})$ = 320 (triệu đồng)

Áp dụng công thức tính độ lệch chuẩn của $\overline{X}$:

$\sigma_{\overline{X}}^2 = \sum(x - \overline{X})^2.P(\overline{X})$ = 975 (triệu đồng)^2

$\sigma_{\overline{X}} = \sqrt{\sigma_{\overline{X}}^2}$ = 31,22499 (triệu đồng)

Ý nghĩa của các trị số này:

$\mu_{\overline{X}}$ cho biết giá trị trung bình của 10 trung bình mẫu đã chọn ở trên, giá trị này cũng bằng với trung bình của tổng thể, $\mu_{\overline{X}}$ là ước lượng không chệch của trung bình tổng thể.

$\sigma_{\overline{X}}$ thể hiện sự biến thiên hay mức độ phân tán của 10 trung bình mẫu xung quanh giá trị kỳ vọng của chúng, có thể nhận thấy các giá trị trung bình mẫu có mức độ phân tán quanh giá trị kì vọng nhỏ hơn so với tổng thể.

Các trị thống kê của mẫu có thể sử dụng để ước lượng các tham số của tổng thể.

d. Chỉ ra và giải thích các mối quan hệ giữa trị thống kê và tham số thống kê nói trên.

Kỳ vọng của phân phối trung bình mẫu bằng kỳ vọng của tổng thể vì kỳ vọng của trung bình mẫu chính là trung bình của các số trung bình của mẫu. Kỳ vọng của phân phối trung bình mẫu là ước lượng không chệch của trung bình tổng thể. Mối liên hệ: $\mu_{\overline{X}} = \mu$

Độ lệch chuẩn của phân phối trung bình mẫu nhỏ hơn độ lệch chuẩn của tổng thể, vì các trung bình mẫu dao động nhỏ hơn so với các giá trị trong tổng thể. Đối với tổng thể hữu hạn, ta có công thức thể hiện mối liên hệ như sau:

$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{50,99}{\sqrt{2}} \sqrt{\frac{5-2}{5-1}}$ = 31,22499 (triệu đồng)

<u>Bài 2:</u> (30 điểm)

Mục GIADAT trong tập tin dữ liệu *MPP07-521-P04V-Data.xls* trình bày số liệu khảo sát giá đất (triệu đồng/m^2) trên 635 hộ gia đình ở thành phố Hồ Chí Minh. Coi đây là một tổng thể.

a. Tính toán các giá trị thống kê mô tả của tập dữ liệu GIADAT.

b. Chọn ngẫu nhiên một mẫu có cỡ bất kỳ từ tổng thể nêu trên, hãy dự đoán một cách định tính:

- Giá trị kỳ vọng của mẫu này.

- Độ phân tán của giá trị trung bình trong mẫu này xung quanh giá trị trung bình của tổng thể. Nêu rõ các giả định cần thiết để dự đoán của bạn có độ tin cậy cao.

c. Hãy chọn ngẫu nhiên một mẫu có cỡ n = 75 từ tập dữ liệu nói trên, chỉ ra các giá trị:

- Kỳ vọng của mẫu.
- Khoảng cách từ kỳ vọng của mẫu tới kỳ vọng của tổng thể.
- So sánh với dự đoán ở phần (b) và giải thích các quan sát thu được.

d. Dùng lệnh Random trong Excel để chọn ra 25 mẫu ngẫu nhiên bất kỳ từ tập dữ liệu trên với cỡ mẫu tăng dần (theo cấp số cộng, công sai là 5) bắt đầu từ n = 150.

- Với mỗi mẫu hãy lặp lại các yêu cầu trong phần (c).
- Vẽ đồ thị của các giá trị vừa tìm được theo sự tăng dần của n.
- Mô tả xu thế thay đổi của các giá trị này theo n. Bạn rút ra được điều gì từ câu hỏi này?

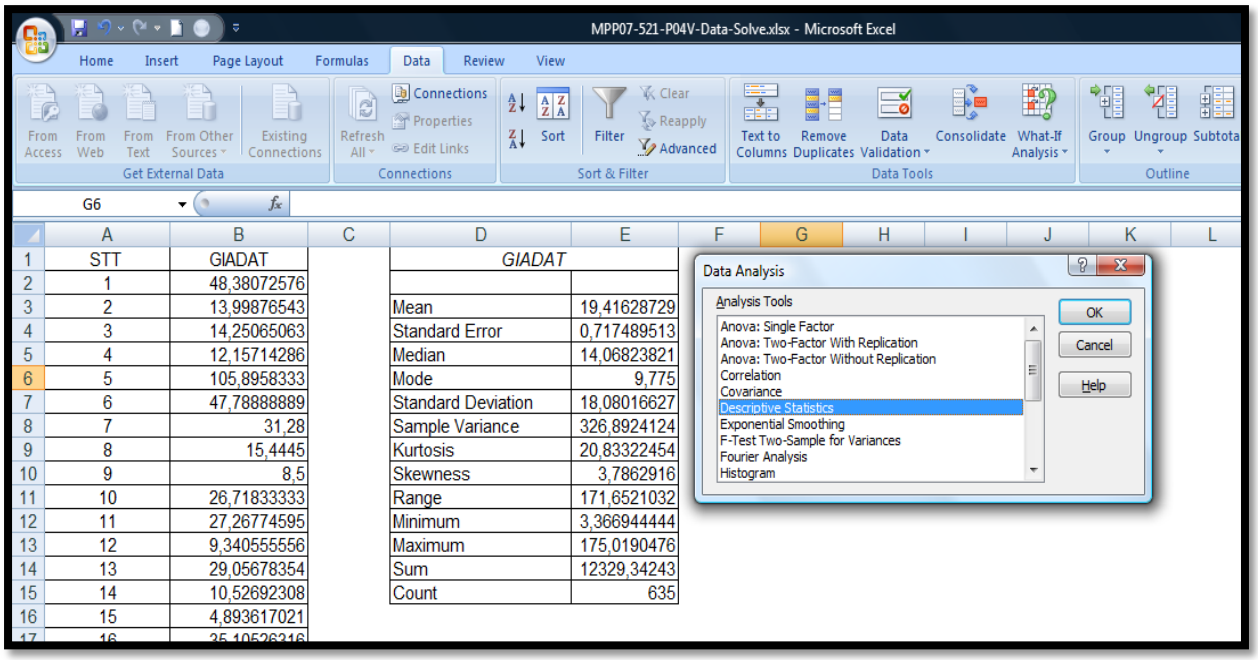
e. Giả sử rằng, số quan sát của biến GIADAT rất lớn (lớn hơn rất nhiều so với 635 quan sát) các giá trị trung bình và phương sai được cho là không đổi so với các giá trị trong phần (a). Với một mong muốn có độ lệch chuẩn của mẫu chọn từ tập dữ liệu này là 45 nghìn đồng/m^2, thì cỡ mẫu cần thiết là bao nhiêu?

Đáp án bài 2:

a. Tính toán các giá trị thống kê mô tả của tập dữ liệu GIADAT.

Trong Ms Excel 2007, ta sử dụng chức năng Descriptive Statistic để tính toán các giá trị thống kê mô tả của biến GIADAT như sau:

Chọn Data -> Data Analysis -> Descriptive Statistic và chọn như hình dưới:



STT	GIADAT
1	48,38072576
2	13,99876543
3	14,25065063
4	12,15714286
5	105,8958333
6	47,78888889
7	31,28
8	15,4445
9	8,5
10	26,71833333
11	27,26774595
12	9,340555556
13	29,05678354
14	10,52692308
15	4,893617021

GIADAT	
Mean	19,41628729
Standard Error	0,717489513
Median	14,06823821
Mode	9,775
Standard Deviation	18,08016627
Sample Variance	326,8924124
Kurtosis	20,83322454
Skewness	3,7862916
Range	171,6521032
Minimum	3,366944444
Maximum	175,0190476
Sum	12329,34243
Count	635

Kết quả thu được:

GIADAT		
Mean	19,41628729	(triệu đồng/m^2)
Standard Error	0,717489513	(triệu đồng/m^2)
Median	14,06823821	(triệu đồng/m^2)
Mode	9,775	(triệu đồng/m^2)
Standard Deviation	18,08016627	(triệu đồng/m^2)
Sample Variance	326,8924124	(triệu đồng/m^2)^2
Kurtosis	20,83322454	
Skewness	3,7862916	
Range	171,6521032	(triệu đồng/m^2)
Minimum	3,366944444	(triệu đồng/m^2)
Maximum	175,0190476	(triệu đồng/m^2)
Sum	12329,34243	(triệu đồng/m^2)
Count	635	

Lưu ý: Với các tính này, Excel coi tập dữ liệu là một mẫu thay vì tổng thể, do đó các giá trị Phương sai và Độ lệch chuẩn được tính theo công thức của mẫu.

Nếu coi tập dữ liệu là tổng thể thì dùng lệnh =Varp(GIADAT) của excel ta có:

σ^2= 326,377621 (triệu đồng/m^2)^2

⇨ $\sigma = \sqrt{326,377621}$ = 18,065924 (triệu đồng/m^2)

b. Chọn ngẫu nhiên một mẫu có cỡ bất kỳ từ tổng thể nêu trên, hãy dự đoán một cách định tính:

- Giá trị kỳ vọng của mẫu này.
- Độ phân tán của giá trị trung bình trong mẫu này xung quanh giá trị trung bình của tổng thể. Nêu rõ các giả định cần thiết để dự đoán của bạn có độ tin cậy cao.

Tự chọn ngẫu nhiên một mẫu với cỡ mẫu bất kì, giả sử cỡ mẫu lớn hơn 30. Theo Định lý giới hạn Trung tâm, nếu các mẫu ngẫu nhiên gồm n quan sát được rút ra từ một tổng thể không chuẩn tắc với trung bình μ và độ lệch chuẩn σ, thì khi n lớn, phân phối chọn mẫu của trung bình mẫu $\bar{x}$ được phân phối xấp xỉ chuẩn tắc với $\mu_{\bar{x}} = \mu$ và $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Trong trường hợp này, vì tổng thể hữu hạn gồm 635 quan sát nên công thức $\sigma_{\bar{x}}$ phải có hệ số điều chỉnh như sau: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Vậy có thể dự đoán:

- Giá trị kỳ vọng của mẫu phân tán xung quanh giá trị kỳ vọng của tổng thể. Nếu phải chọn một giá trị cụ thể thì giá trị trung bình của tổng thể sẽ là con số để lựa chọn tốt nhất trong trường hợp này (chúng ta hi vọng mẫu sẽ đại diện được cho tổng thể).

- Theo Định lý giới hạn Trung tâm, phân phối chọn mẫu của trung bình mẫu $\bar{x}$ được phân phối xấp xỉ chuẩn tắc với $\mu_{\bar{x}}$ và $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ nên xác suất là 95% các trung bình mẫu sẽ nằm trong giới hạn hai độ lệch chuẩn tính từ trung bình của trung bình mẫu. Vậy độ phân tán của giá trị trung bình trong mẫu này xung quanh giá trị trung bình của tổng thể (hay khoảng cách từ giá trị trung bình trong mẫu này đến giá trị trung bình của tổng thể) sẽ xoay quanh độ lệch chuẩn của phân phối trung bình mẫu với xác suất 95% nằm trong khoảng (-2$\sigma_{\bar{x}}$; 2$\sigma_{\bar{x}}$)

Để thuận tiện cho việc so sánh với câu c, ta chọn mẫu ngẫu nhiên với cỡ mẫu n = 75 từ tổng thể trên. Có thể dự đoán một cách định tính:

- Giá trị kì vọng của mẫu xấp xỉ 19,416287 (triệu đồng/m^2)

- Độ phân tán của giá trị trung bình trong mẫu này xung quanh giá trị trung bình của tổng thể sẽ xoay quanh $\sigma_{\bar{x}}$.

Với $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{18{,}065924}{\sqrt{75}}\sqrt{\frac{635-75}{635-1}}$ = 1,96055 (triệu đồng/m^2)

Với xác suất 95%, giá trị tuyệt đối của chênh lệch từ giá trị trung bình trong mẫu này đến giá trị trung bình của tổng thể sẽ nhỏ hơn hai độ lệch chuẩn hay nhỏ hơn: 2*1,96055 = 3,9211 (triệu đồng/m^2)

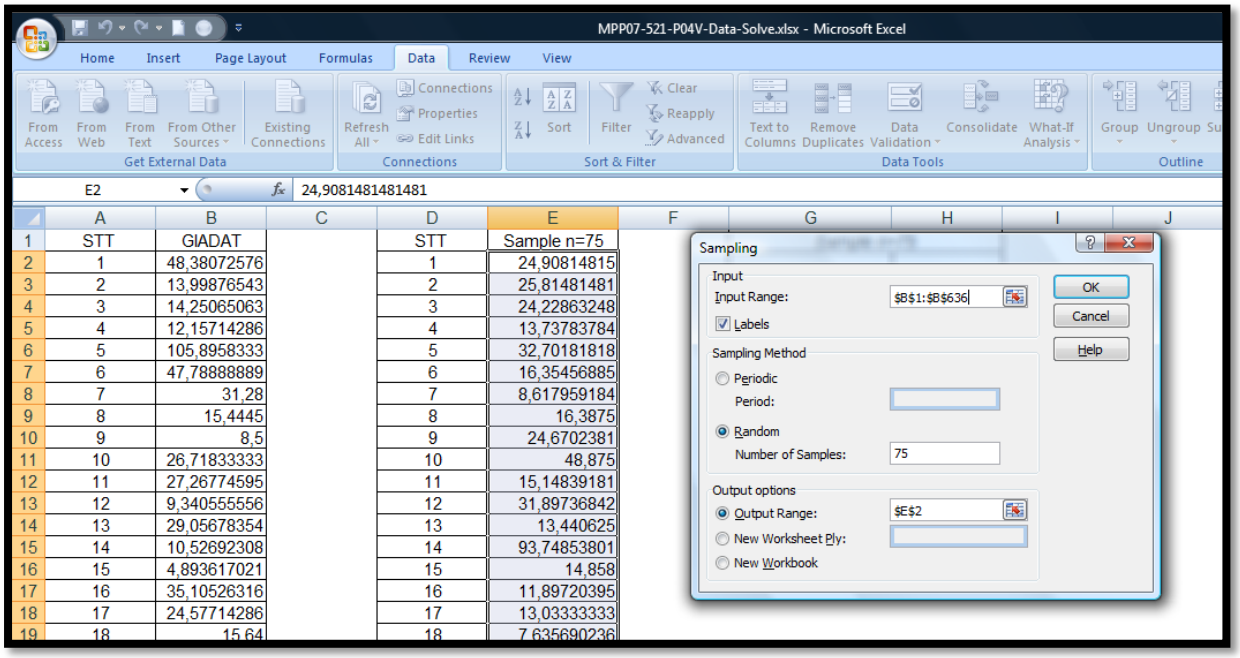
Để có thể dự đoán các giá trị như trên, cần giả định:

- Việc chọn mẫu là ngẫu nhiên
- Kích thước mẫu n là lớn (n>=30)

c. Hãy chọn ngẫu nhiên một mẫu có cỡ n = 75 từ tập dữ liệu nói trên, chỉ ra các giá trị:

- Kỳ vọng của mẫu.
- Khoảng cách từ kỳ vọng của mẫu tới kỳ vọng của tổng thể.
- So sánh với dự đoán ở phần (b) và giải thích các quan sát thu được.

Ta tiếp tục chọn ngẫu nhiên một mẫu có cỡ n = 75 như sau:



	A	B	C	D	E
1	STT	GIADAT		STT	Sample n=75
2	1	48,38072576		1	24,90814815
3	2	13,99876543		2	25,81481481
4	3	14,25065063		3	24,22863248
5	4	12,15714286		4	13,73783784
6	5	105,8958333		5	32,70181818
7	6	47,78888889		6	16,35456885
8	7	31,28		7	8,617959184
9	8	15,4445		8	16,3875
10	9	8,5		9	24,6702381
11	10	26,71833333		10	48,875
12	11	27,26774595		11	15,14839181
13	12	9,340555556		12	31,89736842
14	13	29,05678354		13	13,440625
15	14	10,52692308		14	93,74853801
16	15	4,893617021		15	14,858
17	16	35,10526316		16	11,89720395
18	17	24,57714286		17	13,03333333
19	18	15,64		18	7,635690236

- Kỳ vọng của mẫu:

Sử dụng lệnh Average, ta tính được kỳ vọng của mẫu là: 22,259352 (triệu đồng/m^2)

- Khoảng cách từ kỳ vọng của mẫu tới kỳ vọng của tổng thể:

Khoảng cách (hay chênh lệch) giữa kỳ vọng của mẫu tới kỳ vọng của tổng thể là:

22,259352 – 19,416287 = 2,843064 (triệu đồng/m^2)

- So sánh với dự đoán ở phần (b) và giải thích các quan sát thu được.

So sánh với kết quả ở câu (b) ta thấy:

- Trung bình của mẫu vừa chọn là 22,259352 có chênh lệch không lớn với kỳ vọng của trung bình mẫu là 19,416287.

- Mức chênh lệch 2,843064 cũng xoay quanh độ lệch chuẩn của phân phối trung bình mẫu và mức chênh lệch 2,843064 cũng nhỏ hơn hai độ lệch chuẩn của phân phối trung bình mẫu như đã dự đoán (2,843064 < 3,9211).

Giải thích:

- Với mẫu vừa chọn ở câu (c), giá trị trung bình của mẫu là 22,259352 và là một đại diện trong phân phối trung bình mẫu mà ta đề cập tới ở câu (b), nên giá trị trung bình này (22,259352) là một phần tử xoay quanh giá trị trung bình của phân phối mẫu (19,416287).

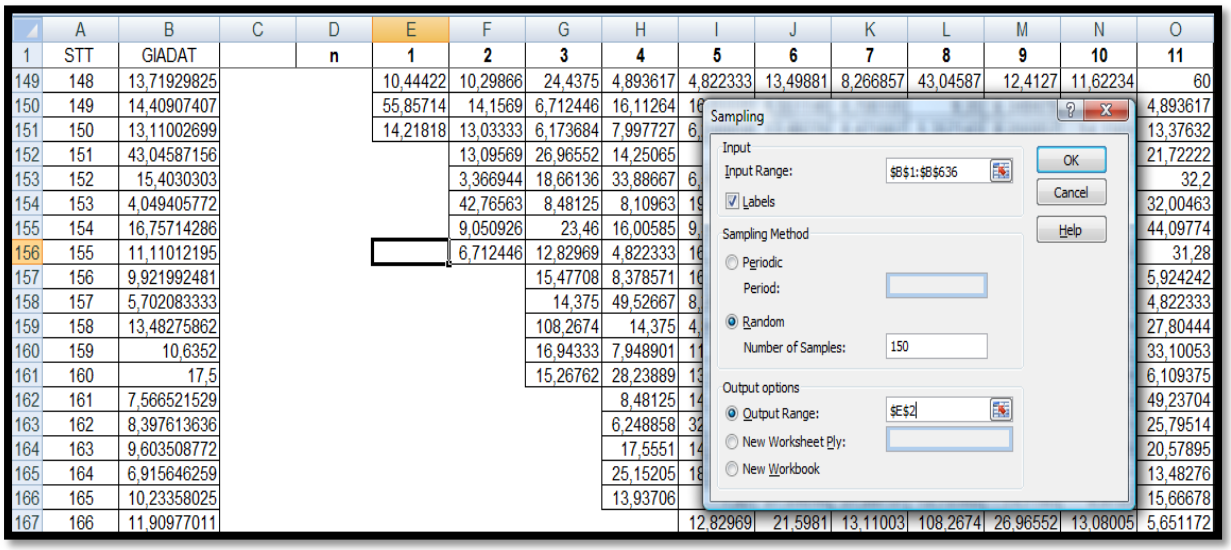
- Khoảng cách giữa kỳ vọng của mẫu tới kỳ vọng của tổng thể (cũng bằng kỳ vọng của phân phối trung bình mẫu) thể hiện mức độ phân tán của kỳ vọng của mẫu so với kỳ vọng của tổng thể, nên giá trị của nó cũng xoay quanh độ lệch chuẩn của phân phối trung bình mẫu nằm trong giới hạn hai độ lệch chuẩn với xác suất 95% (vì phân phối trung bình mẫu xấp xỉ chuẩn tắc).

d. Dùng lệnh Random trong Excel để chọn ra 25 mẫu ngẫu nhiên bất kỳ từ tập dữ liệu trên với cỡ mẫu tăng dần (theo cấp số cộng, công sai là 5) bắt đầu từ n = 150.

- Với mỗi mẫu hãy lặp lại các yêu cầu trong phần (c).
- Vẽ đồ thị của các giá trị vừa tìm được theo sự tăng dần của n.
- Mô tả xu thế thay đổi của các giá trị này theo n. Bạn rút ra được điều gì từ câu hỏi này?

Để chọn ra 25 mẫu ngẫu nhiên bất kỳ từ tập dữ liệu trên với cỡ mẫu tăng dần bắt đầu từ n = 150, ta làm lần lượt 25 lần như sau:

Chọn Data -> Data Analysis -> Sampling và điền các thông số như sau:



Làm lần lượt như vậy 25 lần, mỗi lần thay đổi tăng số Number of Samples trong mục Random lên 5 đơn vị từ cỡ mẫu n = 150 cho đến khi n = 270, ta được 25 mẫu ngẫu nhiên bất kì thỏa mãn đề bài.

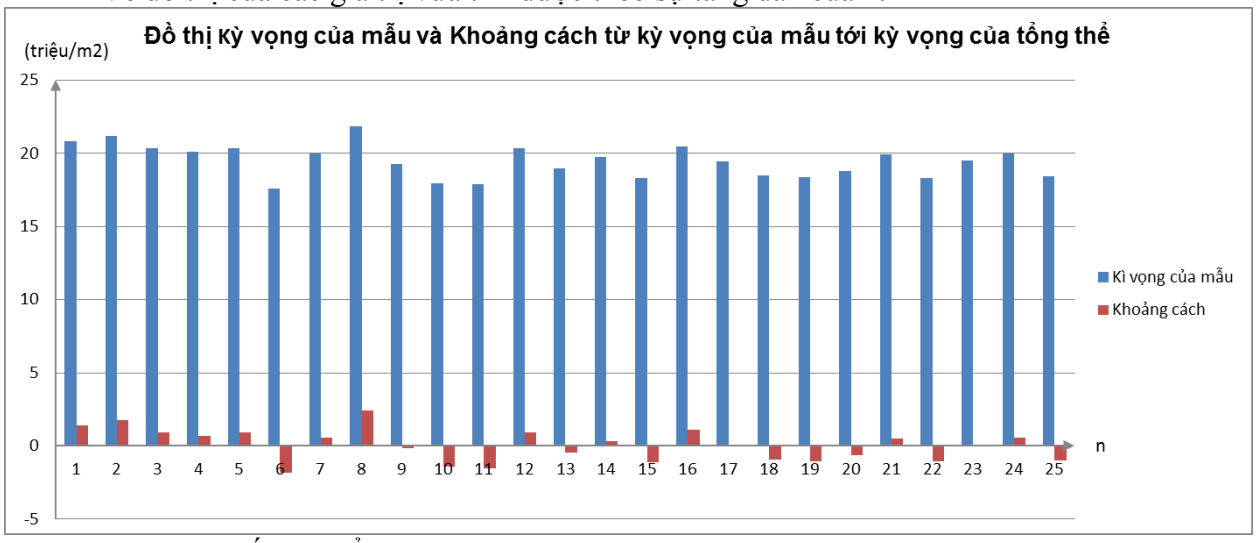
- Với mỗi mẫu hãy lặp lại các yêu cầu trong phần (c).

Với mỗi mẫu, tính kỳ vọng của mẫu; khoảng cách từ kỳ vọng của mẫu tới kỳ vọng của tổng thể như sau:

	A	B	C	D	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	STT	GIADAT		n	14	15	16	17	18	19	20	21	22	23	24	25
255	254	7,680357143											10,28947	4,822333	14,1569	23,17037
256	255	6,873046875											14,81288	4,444286	25,46047	15,26762
257	256	20,332												42,76563	14,15048	22,00433
258	257	16,94333333												12,03077	13,93706	7,680357
259	258	16,57748538												8,460234	10,47321	20,57895
260	259	89,37142857												14,41813	14,1569	4,8875
261	260	5,357142857												12,28857	99,36126	5,674376
262	261	44,09774436													17,10625	8,960417
263	262	6,187542088													40,72917	24,22863
264	263	9,775													5,702083	16,03335
265	264	66,21212121													41,32939	13,62369
266	265	28,30095238													42,76563	15,82619
267	266	15,34378788														17,9501
268	267	13,65625														4,265455
269	268	16,3875														17,83509
270	269	49,23703704														39,89796
271	270	14,81288462														18,15357
272	271	4,25		Kì vọng của mẫu	19,7659	18,33782	20,49922	19,46215	18,50565	18,37667	18,79707	19,94467	18,34019	19,51521	19,96792	18,44458
273	272	26,15686275		Khoảng cách	0,349612	-1,07846	1,082937	0,045862	-0,91064	-1,03962	-0,61922	0,528386	-1,07609	0,098923	0,551692	-0,97171
274	273	10,47321429														
275	274	5,5														
276	275	20,36458333														
277	276	8,759612403														

=AVERAGE(AC2:AC271)
=AVERAGE(AC2:AC271) - 19,4162872926061

- Vẽ đồ thị của các giá trị vừa tìm được theo sự tăng dần của n.



- Mô tả xu thế thay đổi của các giá trị này theo n.

Khi n tăng lên:

- *Kỳ vọng của mẫu* vẫn biến động xung quanh giá trị kỳ vọng của tổng thể (19,416287).

- *Khoảng cách từ kỳ vọng của mẫu tới kỳ vọng của tổng thể* cũng dao động xung quanh giá trị 0 và có vẻ như ngày càng dao động nhỏ hơn khi n càng tăng.

- Bạn rút ra được điều gì từ câu hỏi này?

Như vậy, nếu chọn ngẫu nhiên cỡ mẫu càng lớn, thì sự biến động của *Kỳ vọng của mẫu* thể hiện bằng *Khoảng cách từ kỳ vọng của mẫu tới kỳ vọng của tổng thể* càng nhỏ, độ chính xác của trung bình mẫu so với kỳ vọng của tổng thể càng cao. Tuy nhiên việc gia tăng cỡ mẫu sẽ phát sinh

thêm thời gian và chi phí thực hiện. Do đó cần cân nhắc để đánh đổi hợp lí giữa độ chính xác mong muốn và chi phí phải bỏ ra.

(Đối với các trường hợp mà dao động của *Khoảng cách từ kỳ vọng của mẫu tới kỳ vọng của tổng thể* không có xu hướng giảm khi n tăng thì cần cân nhắc lại việc lấy mẫu. Mẫu được chọn có thể không mang tính đại diện cho tổng thể. Kết luận đưa ra là việc chọn cỡ mẫu tăng dần tuy nhìn chung sẽ gia tăng được độ chính xác, nhưng nếu các phần tử trong mẫu thiên lệch thì kết quả nhận được vẫn chênh lệch nhiều so với tổng thể. Vì vậy việc đầu tư cho cỡ mẫu lớn có thể không phát huy tác dụng. Cần phải điều chỉnh cách lấy mẫu cho phù hợp).

e. Giả sử rằng, số quan sát của biến GIADAT rất lớn (lớn hơn rất nhiều so với 635 quan sát) các giá trị trung bình và phương sai được cho là không đổi so với các giá trị trong phần (a). Với một mong muốn có độ lệch chuẩn của mẫu chọn từ tập dữ liệu này là 45 nghìn đồng/m^2, thì cỡ mẫu cần thiết là bao nhiêu?

Lúc này 635 quan sát có được là một mẫu của tổng thể. Vì giả định các giá trị trung bình và phương sai được cho là không đổi so với các giá trị trong phần (a), nên ta có Trung bình tổng thể = 19,416287 (triệu đồng/m^2), và Phương sai tổng thể = 326,377621 (triệu đồng/m^2)^2.

Với mong muốn có độ lệch chuẩn của mẫu chọn từ tập dữ liệu này là 45 nghìn đồng/m^2 hay 0,045 (triệu đồng/m^2), ta có:

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ => n = $\frac{\sigma^2}{\sigma_{\bar{x}}^2} = \frac{326{,}377621}{0{,}045^2}$ = 161.174,13 ≈ 161.175 (hộ gia đình)

⇨ Cỡ mẫu cần thiết là n = 161.175 (hộ gia đình)

(Theo nguyên tắc thận trọng, để có độ chính xác tối thiểu mong muốn thì số lượng mẫu điều tra luôn phải được làm tròn lên)

<u>Bài 3:</u> (15 điểm)

Từ một mẫu tổng quát W = (X_1, X_2), giả sử phương sai của X_1, X_2 là như nhau, chúng ta xem xét ước lượng trung bình của tổng thể μ như sau:

$\overline{X} = aX_1 + bX_2$, với (a + b) = 1, a > 0, b > 0

a. Chứng minh rằng $\overline{X}$ là ước lượng không chệch của μ.

b. Với điều kiện nào của a, b thì ước lượng $\overline{X}$ hiệu quả nhất?

<u>Đáp án bài 3:</u>

a. Ta có $E(\overline{X}) = E(aX_1 + bX_2) = E(aX_1) + E(bX_2) = aE(X_1) + bE(X_2)$

$= a.\mu + b.\mu = (a + b)\mu = \mu$

Vậy $\overline{X}$ là ước lượng không chệch của μ.

b. Ta có $Var(\overline{X}) = Var(aX_1 + bX_2) = Var(aX_1) + Var(bX_2) = a^2.Var(X_1) + b^2.Var(X_2)$

$= a^2.\sigma^2 + b^2.\sigma^2 = (a^2 + b^2)\sigma^2$

Để ước lượng $\overline{X}$ hiệu quả nhất thì $Var(\overline{X})$ phải nhỏ nhất.

Tìm điều kiện để $Var(\overline{X})$ min ⇔ $(a^2 + b^2)\sigma^2$ min ⇔ $a^2 + b^2$ min

Đặt $F = a^2 + b^2 = a^2 + (1 - a)^2 = 2a^2 - 2a + 1$

Đạo hàm bậc 1 theo a: $F' = d(2a^2 - 2a + 1)/da = 4a - 2$

Đạo hàm bậc 2 theo a: F'' = d(4a – 2)/da = 4 > 0

Để F min ⇔ F' = 4a – 2 = 0 ⇔ a = ½

Điều kiện để Var($\overline{X}$) min là a = b = ½

(Có thể sử dụng các phép biến đổi đa thức, bất đẳng thức Cauchy, Bunyakovsky… để giải)

Bài 4: (20 điểm)

Chọn ngẫu nhiên 240 người dân trong khu phố để khảo sát thì thấy 160 người có tham gia bảo hiểm y tế (BHYT). Hãy trả lời các câu hỏi sau, nêu rõ cách tính toán và các lập luận:

a. Ước lượng sai số chuẩn của phân phối mẫu.

b. Với độ tin cậy 95%, tỉ lệ người dân trong khu phố tham gia BHYT là bao nhiêu?

c. Giả định tỉ lệ người dân trong khu phố tham gia BHYT trên thực tế là 70%. Xác suất để tỉ lệ người dân tham gia BHYT trong mẫu thấp hơn 60% là bao nhiêu?

d. Theo thông tin câu (c), để độ chính xác là 5% và độ tin cậy là 95% cần khảo sát bao nhiêu người dân trong khu phố?

Đáp án bài 4:

n = 240, x = 160 => $\hat{p}$ = 160/240 = 66,67%

a. Ước lượng sai số chuẩn của phân phối mẫu.

Sai số chuẩn (độ lệch chuẩn) của phân phối mẫu :

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{66{,}67\%(1-66{,}67\%)}{240}} = 0{,}030429031 \approx 3{,}04\%$$

b. Với độ tin cậy 95%, tỉ lệ người dân trong khu phố tham gia BHYT là bao nhiêu?

Độ tin cậy 95% => $\alpha = 1 - 95\% = 5\%$ => $Z_{\alpha/2} = 1{,}96$

Biên sai số là : $Z_{\alpha/2}.\sigma_{\hat{p}} = 1{,}96*3{,}04\% = 5{,}96\%$

Khoảng tin cậy ước lượng cho tỉ lệ tham gia BHYT của người dân trong khu phố là :

$$\hat{p} \pm Z_{\alpha/2}.\sigma_{\hat{p}}$$

Thay vào công thức ta được : $66{,}67\% \pm 5{,}96\%$

Vậy ta có khoảng tỉ lệ (60,7% ; 72,6%) là khoảng ước lượng chứa tỉ lệ tham gia BHYT thực sự của người dân trong khu phố với độ tin cậy 95%.

c. Giả định tỉ lệ người dân trong khu phố tham gia BHYT trên thực tế là 70%. Xác suất để tỉ lệ người dân tham gia BHYT trong mẫu thấp hơn 60% là bao nhiêu?

p = 70%, $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{70\%(1-70\%)}{240}} = 0{,}029580399 \approx 2{,}96\%$

Xác suất để tỉ lệ người dân tham gia BHYT trong mẫu thấp hơn 60% :

$$Z = \frac{\hat{p}-p}{\sigma_{\hat{p}}} = \frac{60\%-70\%}{2{,}96\%} = -3{,}38$$

Dùng hàm Normsdist trên excel ta có:

P($\hat{p}$ < 60%) = P ($Z < -3{,}38$) = 0,000361616 ≈ 0,036%

d. Theo thông tin câu (c), để độ chính xác là 5% và độ tin cậy là 95% cần khảo sát bao nhiêu người dân trong khu phố?

Với độ chính xác là 5% và độ tin cậy là 95% ta có :

$Z_{\alpha/2}.\sigma_{\hat{p}} = 1{,}96*\sqrt{\frac{pq}{n}} = 1{,}96*\sqrt{\frac{70\%(1-70\%)}{n}} = 5\%$ => n = 322,6944 (người dân)

Vậy cần khảo sát 323 người dân.

<u>Bài 5:</u> (15 điểm)

Nghiên cứu tìm hiểu nguyên nhân vắng học của học viên với lí do phải chăm sóc con nhỏ, nhà trường đã chọn ra một mẫu ngẫu nhiên gồm 69 học viên có con nhỏ dưới 10 tuổi và thấy rằng thời gian vắng học trung bình là 2,25 giờ mỗi tháng với độ lệch chuẩn 2,09 giờ mỗi tháng. Song song đó nhà trường cũng chọn ra một mẫu ngẫu nhiên độc lập khác gồm 206 học viên không có con nhỏ dưới 10 tuổi và thấy thời gian vắng học trung bình là 1,69 giờ mỗi tháng, độ lệch chuẩn của mẫu là 1,91 giờ mỗi tháng.

a. Xác định khoảng tin cậy 90% và khoảng tin cậy 99% cho sự khác biệt về thời gian vắng học trung bình giữa nhóm học viên có con nhỏ dưới 10 tuổi và nhóm học viên không có con nhỏ dưới 10 tuổi.

b. Giải thích kết quả tính toán.

c. Kết quả này có phải là bằng chứng xác thực giúp chúng ta kết luận rằng học viên có con nhỏ dưới 10 tuổi vắng học nhiều hơn học viên không có con nhỏ dưới 10 tuổi hay không?

<u>Đáp án bài 5:</u>

a. Xác định khoảng tin cậy 90% và khoảng tin cậy 99% cho sự khác biệt về thời gian vắng học trung bình của hai nhóm học viên.

Đối với nhóm học viên có con nhỏ dưới 10 tuổi:

n_1 = 69 (học viên); $\overline{x_1}$ = 2,25 (giờ/tháng); s_1 = 2,09 (giờ/tháng)

Đối với nhóm học viên không có con nhỏ dưới 10 tuổi:

n_2 = 206 (học viên); $\overline{x_2}$ = 1,69 (giờ/tháng); s_2 = 1,91 (giờ/tháng)

Cỡ mẫu n_1, n_2 > 30 nên áp dụng công thức ước lượng khoảng tin cậy cho sự khác biệt về thời gian vắng học trung bình của hai nhóm học viên:

$$(\overline{x_1} - \overline{x_2}) - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

- Với độ tin cậy 90% => $\alpha = 1 - 90\% = 10\%$ => $Z_{\alpha/2} = 1{,}645$

Khoảng tin cậy 90%:

$$(2{,}25 - 1{,}69) - 1{,}645\sqrt{\frac{2{,}09^2}{69}+\frac{1{,}91^2}{206}} < \mu_1 - \mu_2 < (2{,}25 - 1{,}69) + 1{,}645\sqrt{\frac{2{,}09^2}{69}+\frac{1{,}91^2}{206}}$$

Hay: 0,09 < $\mu_1 - \mu_2$ < 1,03 (giờ/tháng)

- Với độ tin cậy 99% => $\alpha = 1 - 99\% = 1\%$ => $Z_{\alpha/2} = 2{,}576$

Khoảng tin cậy 99%:

$$(2{,}25 - 1{,}69) - 2{,}576\sqrt{\frac{2{,}09^2}{69} + \frac{1{,}91^2}{206}} < \mu_1 - \mu_2 < (2{,}25 - 1{,}69) + 2{,}576\sqrt{\frac{2{,}09^2}{69} + \frac{1{,}91^2}{206}}$$

Hay: $-0{,}17 < \mu_1 - \mu_2 < 1{,}29$ (giờ/tháng)

b. Giải thích kết quả tính toán.

- Khoảng thời gian từ 0,09 giờ/tháng cho đến 1,03 giờ/tháng là khoảng thời gian ước lượng chứa sự khác biệt thực tế về thời gian vắng học trung bình giữa nhóm học viên có con nhỏ dưới 10 tuổi và nhóm học viên không có con nhỏ dưới 10 tuổi với độ tin cậy 90%.

- Khoảng thời gian từ -0,17 giờ/tháng cho đến 1,29 giờ/tháng là khoảng thời gian ước lượng chứa sự khác biệt thực tế về thời gian vắng học trung bình giữa nhóm học viên có con nhỏ dưới 10 tuổi và nhóm học viên không có con nhỏ dưới 10 tuổi với độ tin cậy 99%.

- Kết quả ước lượng từ hai trường hợp trên cho thấy có sự đánh đổi giữa độ tin cậy mong muốn với độ hẹp của khoảng ước lượng. Độ tin cậy cao hơn dẫn đến khoảng tin cậy rộng hơn. khoảng tin cây rộng sẽ an toàn hơn cho kết luận về giá trị trung bình (rủi ro sai là thấp). Tuy nhiên, ở góc độ của người ra quyết định thì khoảng tin cậy hẹp hơn sẽ dễ ra quyết định hơn.

c. Kết quả này có phải là bằng chứng xác thực giúp chúng ta kết luận rằng học viên có con nhỏ dưới 10 tuổi vắng học nhiều hơn học viên không có con nhỏ dưới 10 tuổi hay không?

- Kết quả tính toán về khoảng tin cậy 90% cho phép chúng ta kết luận rằng học viên có con nhỏ dưới 10 tuổi vắng học nhiều hơn học viên không có con nhỏ dưới 10 tuổi vì khoảng tin cậy này chỉ chứa các giá trị dương (>0).

- Kết quả tính toán về khoảng tin cậy 99% không cho phép chúng ta kết luận rằng học viên có con nhỏ dưới 10 tuổi vắng học nhiều hơn học viên không có con nhỏ dưới 10 tuổi vì khoảng tin cậy này chứa đồng thời các giá trị âm, dương và số 0.