Journal of educational equipment: Applied research, Volume 1, Issue 318 (August 2024)
ISSN 1859 - 0810

Sử dụng Data analysis trong Excel để giảng dạy Ước lượng tham số và Kiểm định giả thiết thống kê ở môn Xác suất thống kê

Nguyễn Thành Tâm

ThS.Trường Cao đẳng Cộng đồng Đồng Tháp

Received: 9/5/2024; Accepted: 16/5/2024; Published:20/5/2024

Abstract: *This article introduces the Data Analysis Toolkit in MS-Excel to teach statistical probability. Instructions for students to analysis data by themselves, to perform the content of estimating and testing statistical hypotheses. Initially, it helps learners to access statistical software that is simple, easy to use and can be used for future work.*

Keywords: *Data Analysis, Excel, probability statistics, estimates, accreditation.*

1. Đặt vấn đề.

MS - EXCEL thuộc bộ Microsoft Office, đơn giản dễ sử dụng và có khả năng phân tích TK gần như chuyên nghiệp. Vì vậy vận dụng kết hợp công cụ Data Analysis trong Microsoft Excel vào việc giảng dạy phần TK ở môn Xác suất TK thật sự hữu ích và cần thiết.

2. Nội dung nghiên cứu

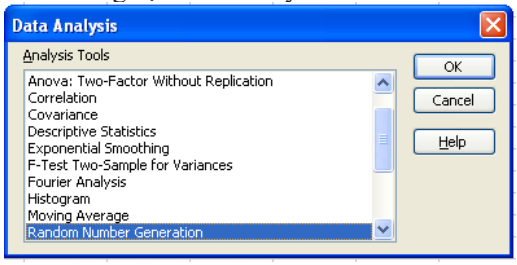
2.1. Thực trạng

Với phần mềm Excel thì đa số sinh viên (SV) đều được tiếp cận và tương đối thành thạo.

2.2.Một số hướng dẫn phân tích TK đơn giản trong excel đến SV

Đối với MS - Excel 2003, công cụ phân tích dữ liệu *Data Analysis* thuộc đơn lệnh *Tools.*

Đối với MS– Excel ở các phiên bản sau không còn tích hợp sẵn trên menu nữa. Để kích hoạt nó ta thực hiện các bước sau: *File - options - Add-Ins -Analysis ToolPak - Go - click vào Analysis ToolPak - ok*. Sau đó vào đơn lệnh Data trong Excel sẽ thấy xuất hiện thanh công cụ Data Analysis



Hộp thoại Data Analysis

2.2.1. Ứng dụng MS- Excel trong việc lấy mẫu.

a) Tạo bảng số ngẫu nhiên.

Nhấp lần lược đơn lệnh *Tools / Data Analysis*

Trong *Data Analysis* chọn *Random Number Generation* rồi nhấn *OK*

Trong *Random Number Generation*, lần lược ấn định các chi tiết sau đây:

Số cột (*Number of Variables*).

Số hàng (*Number of Random Number*),

Loại phân phối (*Uniform, Normal, Bernoulli, Binominal, Poisson, Patterned, Discrete)*

Thông số (*Parameters*)

Mầm ngẫu nhiên *(Random Seed*)

Phạm vi đầu ra (*Output Range*).

b) Lấy một mẫu ngẫu nhiên từ tập hợp tổng thể.

Nhấp *Tools / Data Analysis*

Trong hộp thoại *Data Analysis* chọn Sampling rồi nhấp *OK.*

Trong hộp thoại *Sampling,* lần lược ấn định các chi tiết:

Phạm vi đầu vào (*Input Range*)

Phương pháp (*Sampling Method*): ngẫu nhiên *(Random*) cùng với cở mẫu (*Number of Sample*).

Phạm vi đầu ra (*Output Range*)

c) Tạo các bảng tra phân phối xác suất.

- Tạo bảng phân phối chuẩn tắc với hàm phân phối

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

	A	B	C	D	E	F	G	H
1								
2	z	0.00	0.01	0.02	0.03	...	0,08	0.09
3	0	NORMSDIST($A3+B$2)						
4	0.1							
5	0.2							

Kéo rê cho các hàng và cột còn lại

Tạo bảng hàm Laplace $\varphi(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x}e^{-\frac{t^2}{2}}dt$

Tương tự như trên chỉ khác công thúc: NORMSDIST($A3+B$2) - 0,5

Tạo bảng tra phân vị student P(X>t(n, alpha)) =

alpha.

	A	B	C	D	E	F
1	Bậc tự do	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
2		0.1	0.05	0.025	0.01	0.005
3	1	TINV(B$2,$A3)				
4	2					

Tạo bảng tra phân phối Fisher

Mức ý nghĩa 0,05

	A	B	C	D	E	F	G
1	alpha=0,05						
2	**Bậc tự do v1** →	1	2	3	4	5	…
3	**Bậc tự do v2**						
4	1	FINV(0.05,B$2,$A4)					
5	2						
6	3						

Mức ý nghĩa 0,01 ta làm tương tự thay 0,05 bằng 0,01

2.2.2. Mô tả bộ số liệu và kiểm định giả thiết.

a) Bảng phân phối tần số - tần suất

- Nhập dữ liệu (DL)
- Dùng hàm frequency *(data_array, bins_array)*
- *data_array:* địa chỉ mảng DL
- *bins_array:* Địa chỉ mảng các giá trị khác nhau của DL

VD: Lập bảng và vẽ biểu đồ DL: 12, 13, 11, 13, 15, 12, 11, 10, 14, 13, 12, 15.

Nhập các giá trị, nhập cột các giá trị khác nhau

Đánh dấu khối cột chứa dữ liệu tần số, nhấn F2

Nhập công thức "=frequency(data-array,bins-array), ấn CTRL + SHIFT + ENTER

Lập bảng phân phối tần suất Tần số / tổng số rồi copy cho các ô còn lại.

Vẽ biểu đồ tần suất:

Chọn menu: *insert/Chart/Line .../Next*

Nhập vào *Data Range*: khối DL tần suất và chọn mục *Column*

Chọn *Tab Series*, nhập địa chỉ cột giá trị x ($C$3:$C$8) vào ***Category (X) axis labels***

Chọn *Next/ Finish*

b) Tính các đặc trưng mẫu.

Vào *Tools/ Data Analysis/ Descriptive Statistics.* Nhập các mục:

Input Range: địa chỉ tuyệt đối chứa DL (phạm vi đầu vào).

Grouped By: Column (số liệu theo cột, Row số liệu theo hàng).

Labels in first Row/Column: nhãn DL (Check vào ô này nếu có nhãn ở dòng đầu)

Confidence Level for Mean: độ tin cậy cho trung bình

K-th Largest: 1(1 số lớn nhất, 2 số lớn nhì)

K-th smallest: 1(số nhỏ nhất, số nhỏ nhì)

Output Range: địa chỉ xuất kết quả (phạm vi đầu ra)

Summary Statistics: Bảng kết quả tóm tắt (đánh dấu check nếu muốn hiện bảng thống kê cơ bản)

c) Ước lượng tham số: Để ước lượng trung bình đám đông ta thực hiện các bước:

Nhập dữ liệu mẫu và xử lí DL mẫu bằng thống kê mô tả.

Tính khoảng ước lượng trung bình theo: giá trị trung bình ± độ chính xác.

d) Kiểm định giả thiết

- Vào ***Tools/ Data Analysis/ Descriptive Statistics***
- Hoặc dùng một số hàm sau: average(*number1, number2, ...*): trung bình mẫu,st dev(*number1, number2, ...*): độ lệch chuẩn, var(*number1, number2, ...)*: phương sai mẫu, tinv *(probability, degrees_freedom)*: trả về giá trị t của phân phối student, tdist (x, *degrees_freedom, tails)*: trả về xác suất của phân phối student, Zα = NORMSINV(1- α/2)

e) So sánh hai trung bình

- So sánh hai trung bình với phương sai đã biết hay mẫu lớn (n>=30)

Dùng menu: ***Tool/ Data Analysis…/ z-test: Two sample for Means***

Phân vị 2 phía $z_{\alpha/2}$ là: ***z Critical two-tail***

Nếu $|z| > z_{\alpha/2}$ thì bác bỏ H_0, chấp nhận H_1 và ngược lại

VD: So sánh tốc độ phản ứng của enzyme thủy phân tinh bột ở 2 nhiệt độ khác nhau qua 10 mẫu thử, biết phương sai tương ứng ở mỗi mức độ nhiệt là 1 và 0.98. Thời gian hoàn tất phản ứng như sau: (đv: giây).

Nhập và xử lý dữ liệu

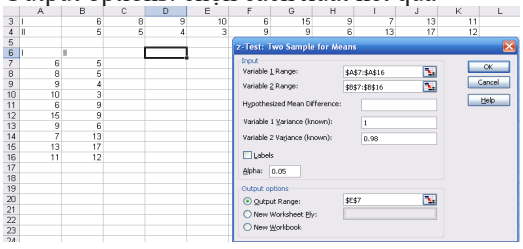
Variable 1 Range, Variable 2 Range: địa chỉ của vùng dữ liệu 40^0C, 50^0C

Variable 1 Variable (known), Variable 2 Variable (known): Phương sai của 40^0C, 50^0C

Labels: Chọn khi có tên biến ở đầu cột hoặc hàng.

Alpha: mức ý nghĩa α

Output options: chọn cách xuất kết quả



- So sánh 2 trung bình với dữ liệu cặp đôi

Chọn menu: ***Tools/ Data Analysis…/ t-test: Paired Two Sample for Means.***

Phân vị 2 phía $t_{\alpha/2}$ là: ***t Critical two-tail***

Nếu $|t| > t_{\alpha/2}$ thì bác bỏ H_0, chấp nhận H_1 và ngược lại

- So sánh hai trung bình với hai phương sai bằng nhau

Nếu dung lượng của cả hai mẫu đều lớn (thường quy ước là $n_1 \geq 30$; $n_2 \geq 30$) ta có thể tiến hành z-test nhưng thay hai phương sai của tổng thể $\sigma_1^2; \sigma_2^2$ bằng hai phương sai mẫu $s_1^2; s_2^2$

Trong trường hợp mẫu bé (n_1, n_2 nhỏ hơn 30) thì ta gặp bài toán khó. Trong trường hợp này, nếu coi hai phương sai của hai tổng thể bằng nhau (cần kiểm định giả thiết phụ về sự bằng nhau của hai phương sai) thì có thể áp dụng tính toán theo phương pháp sau:

Được dùng khi 2 mẫu bé, độc lập và phương sai 2 mẫu bằng nhau.

Chọn menu: ***Tools/Data Analysis…/t-test: Two-Sample Assuming Equal Variances***

Phân vị 2 phía $t_{\alpha/2}$ là: ***t Critical two-tail.***

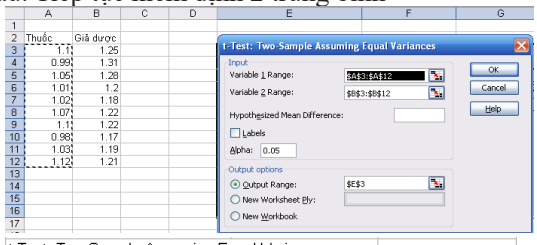
Nếu $|t| > t_{\alpha/2}$ thì bác bỏ H_0, chấp nhận H_1 và ngược lại.

VD: Người ta cho 10 bệnh nhân uống thuốc hạ cholesterol đồng thời cho 10 bệnh nhân khác uống giả dược, sau đó xét nghiệm về nồng độ cholesterol trong máu (g/l) của cả 2 nhóm. Với mức ý nghĩa 0,05 kiểm tra xem thuốc có tác dụng hạ cholesterol trong máu không?

Trước hết ta kiểm tra xem hai phương sai có bằng nhau không

	A	B	C	D	E	F
1	Thuoc	Gia duoc		F-Test Two-Sample for Variances		
2	1.1	1.25				
3	0.99	1.31			*Thuoc*	*Gia duoc*
4	1.05	1.28		Mean	1.047	1.223
5	1.01	1.2		Variance	0.002401	0.002001
6	1.02	1.18		Observations	10	10
7	1.07	1.22		df	9	9
8	1.1	1.22		F	1.199889	
9	0.98	1.17		P(F<=f) one-tail	0.395238	
10	1.03	1.19		F Critical one-tail	3.178893	
11	1.12	1.21				

Ta thấy $F < F_{0,05}$ nên ta xem hai phương sai là bằng nhau. Tiếp tục kiểm định 2 trung bình



t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	*Variable 2*
Mean	1.047	1.223
Variance	0.002401111	0.002001111
Observations	10	10
Pooled Variance	0.002201111	
Hypothesized Mean Difference	0	
df	18	
t Stat	-8.388352782	
P(T<=t) one-tail	6.19807E-08	
t Critical one-tail	1.734063592	
P(T<=t) two-tail	1.23961E-07	
t Critical two-tail	2.100922037	

Kết quả:

$t = -8,3884 < -t_\alpha = -1,7341$ nên chấp nhận H_1. Vậy thuốc trên có tác dụng hạ cholesterol trong máu

Có thể tìm các giá trị theo các hàm Excel:

+ Giá trị P một phía và 2 phía qua hàm TDIST(z,f,1) và TDIST(z,f,2)

+ Giá trị t lý thuyết một phía qua hàm TINV(0.1,f)

+ Giá trị t lý thuyết hai phía qua hàm TINV(0.05,f)

- So sánh hai trung bình với 2 phương sai không bằng nhau

Được dùng khi mẫu bé, độc lập, hai phương sai mẫu không bằng nhau

Chọn menu: ***Tools/ Data Analysis…/ t-test: Two-Sample Assuming unEqual Variances***

Giá trị chuẩn kiểm định: $t = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

Kiểm định 2 phía $t_{\alpha/2}$ là: ***t Critical two-tail***

Nếu $|t| > t_{\alpha/2}$ thì bác bỏ H_0, chấp nhận H_1 và ngược lại

- So sánh hai phương sai.

+ Chọn menu: ***Tools/ Data Analysis…/F-Test Two-Sample for Variances.***

+ Tiêu chuẩn kiểm định $F = \dfrac{S_1^2}{S_2^2}$

+Nếu $F < F_\alpha$ thì chấp nhận $H_0 : \sigma_1^2 = \sigma_2^2$ và ngược lại

Chọn *Tools > Data Analysis > F-Test Two Sample for Variance > OK* sau đó ấn định: Variable 1 Range: miền DL của biến 1 kể cả dòng đầu chứa nhãn.

Variable 2 Range: miền DL của biến 2 kể cả dòng đầu chứa nhãn.

Labels: Chọn mục này nếu miền DL chọn cả dòng nhãn.

Alpha: mức ý nghĩa.

Output Range: Chọn miền trống để đưa ra kết quả.

OK: kết thúc.

3.Kết luận

Hy vọng việc khai thác chức năng Data Analysis vào Xác suất thống kê sẽ giúp cho môn học dễ tiếp cận hơn, việc dạy và học sẽ hiệu quả hơn và bước đầu tạo nền tảng cho việc tiếp cận phân tích số liệu, sẽ có ích cho nhiều người, đặt biệt là những ai về sau cần làm nghiên cứu mà không cần đến những phần mềm thống kê chuyên dụng.

Tài liệu tham khảo

[1] Lê Sĩ Đồng (2007), *Xác suất thống kê và ứng dụng*, NXBGD. Hà Nội

[2] Bộ công cụ Microsoft Excel