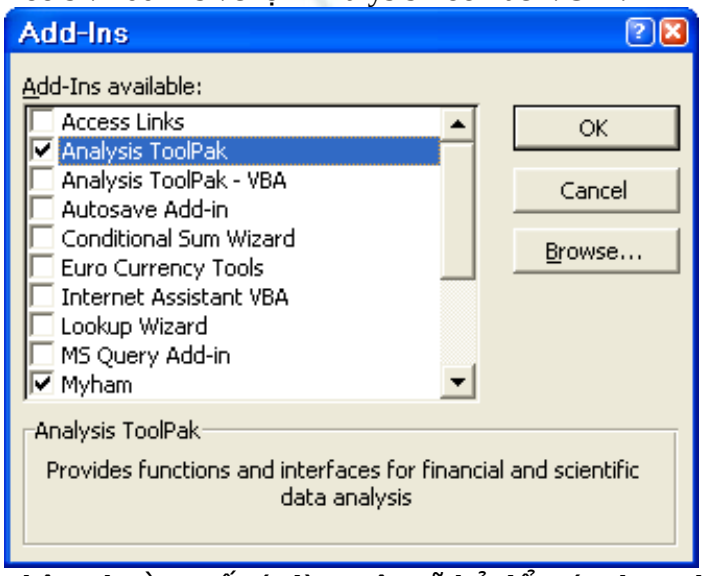
Xử lý thống kê bằng Excel

Module by: Mr Phạm Hữu Duyên

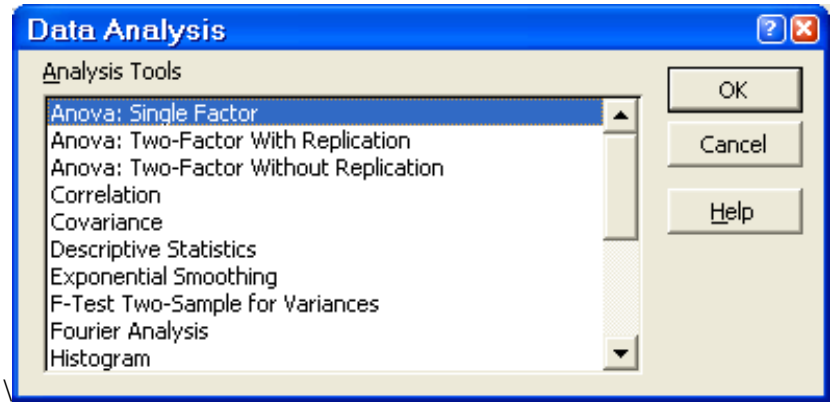
Summary: Dùng Excel để xử lý thống kê với số lượng các mẫu quan sát tương đối nhỏ

Note: Your browser doesn't currently support MathML. If you are using Microsoft Internet Explorer 6 or above, please install the required MathPlayer plugin. Firefox and other Mozilla browsers will display math without plugins, though they require an additional mathematics fonts package. Any browser can view the math in the Print (PDF) version.

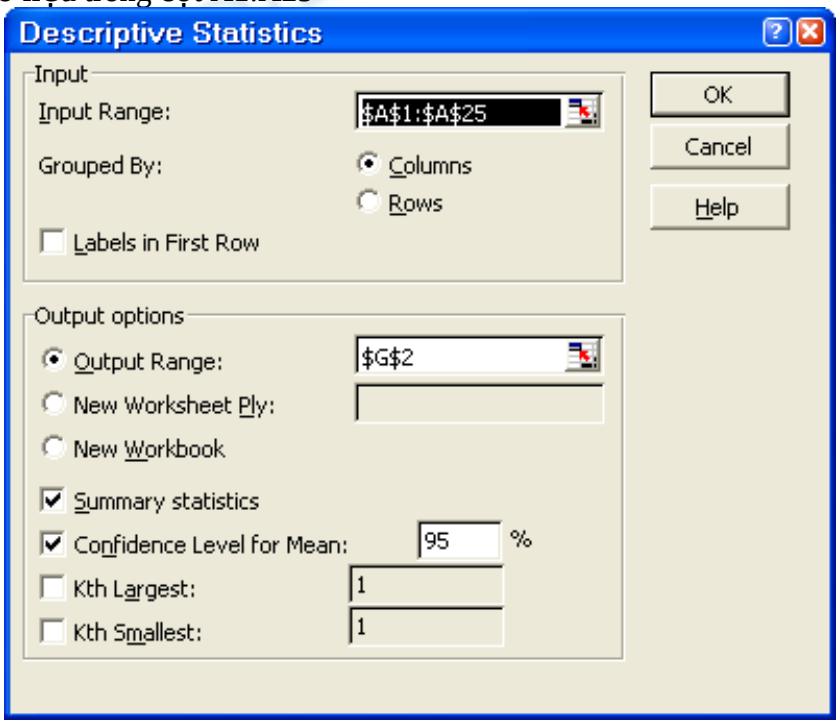
Phân tích số liệu:
Việc phân tích số liệu (xử lý thống kê) có thể được tiến hành bằng các phần mềm chuyên dụng như SPSS, Stat.... Tuy nhiên khi số liệu cần xử lý không nhiều, chủ yếu là các biến định lượng thì có thể sử dụng ngay Analysis ToolPack, một bộ công cụ nhỏ gọn được tích hợp sẵn trong Excel để giải quyết.
Sử dụng Analysis ToolPack.
Nếu trong Tools chưa thấy công cụ này, tiến hành cài đặt theo các bước sau:
Tools \ Add-Ins \ chọn Analysis ToolPack\ OK .



Thông thường nếu ít dùng nên gỡ bỏ để máy chạy nhanh hơn, việc gỡ bỏ ngược lại quá trình cài đặt.



Tools\Data analysis \
Trong bảng chọn có nhiều lựa chọn khác nhau tùy yêu cầu sử dụng.
Xử lý mẫu:
- Sampling: dùng chọn mẫu ngẫu nhiên trong tập hợp khi bố trí thí nghiệm
- Random Number Generation: chọn số ngẫu nhiên tuỳ phương pháp phân phối được lựa chọn, (Uniform, Normal, Bernoulli, Binomial, Poisson, Patterned, Discrete).
- Dùng Descriptive Statistics
Giả sử có số liệu trong cột A1:A25



Hình 1

Kết quả gồm: Kỳ vọng (trung bình), phương sai, trung vị, mode, độ lệch chuẩn, độ nhọn, độ nghiêng (hệ số bất đối xứng so với phân phối chuẩn), khoảng biến thiên, max, min, sum, số mẫu (count), khoảng tin cậy của kỳ vọng ở mức 95%.
Các thông số này có thể được tính theo các hàm tương đương:

Column1			Tính theo hàm
Mean	10,6	Giá trị trung bình	AVERAGE(A1:A25)

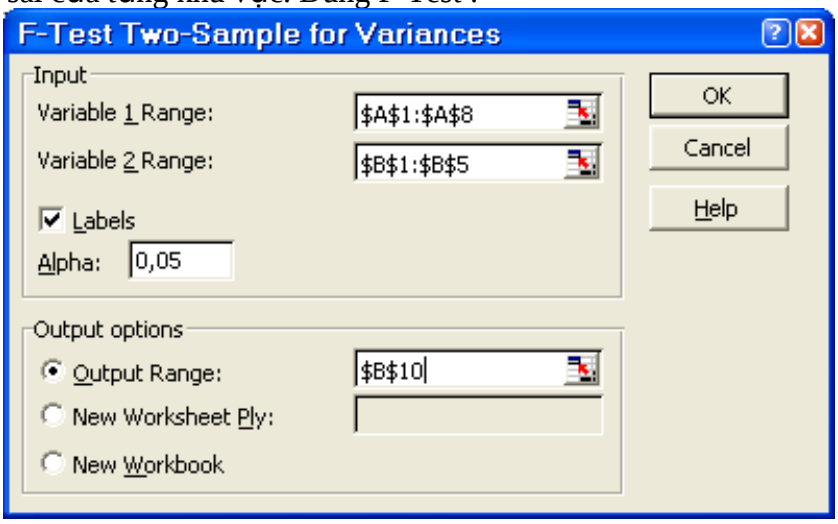
Standard Error	0,41633	Sai số mẫu	
Median	11	Trung vị	MEDIAN(A1:A25)
Mode	11	Mode	MODE(A1:A25)
Standard Deviation	2,08167	Đọ lệch chuẩn	STDEV(A1:A25)
Sample Variance	4,33333	Phương sai mẫu	VAR(A1:A25)
Kurtosis	2,74004	Độ nhọn	KURT(A1:A25)
Skewness	0,91578	Dộ nghiêng	SKEW(A1:A25)
Range	10	Khoảng biến thiên	MAX()-MIN()
Minimum	7	Tối thiểu	MIN(A1:A25)
Maximum	17	Tối đa	MAX(A1:A25)
Sum	265	Tổng	SUM(A1:A25)
Count	25	Số lượng mẫu	COUNT(A1:A25)
Confidence Level(95,0%)	0,85927	Khoảng tin cậy (95,0%)	CONFIDENCE(0,05;I8;I16)

Các kết quả tính toán về thống kê bằng cách dùng Descriptive Statistics và dùng hàm cho kết quả như nhau. Riêng việc xác định khoảng tin cậy (Confidence) cho kết quả khác nhau, do:

- Descriptive Statistics dùng phân bố Student, còn hàm dùng phân bố chuẩn.
- Để thống nhất kết quả cho từng loại phân bố, có thể dùng các hàm khác.

Kiểm định giả thuyết:

- So sánh 2 phương sai: Giả sử có số liệu thí nghiệm của 2 khu vực, so sánh phương sai của từng khu vực. Dùng F-Test :

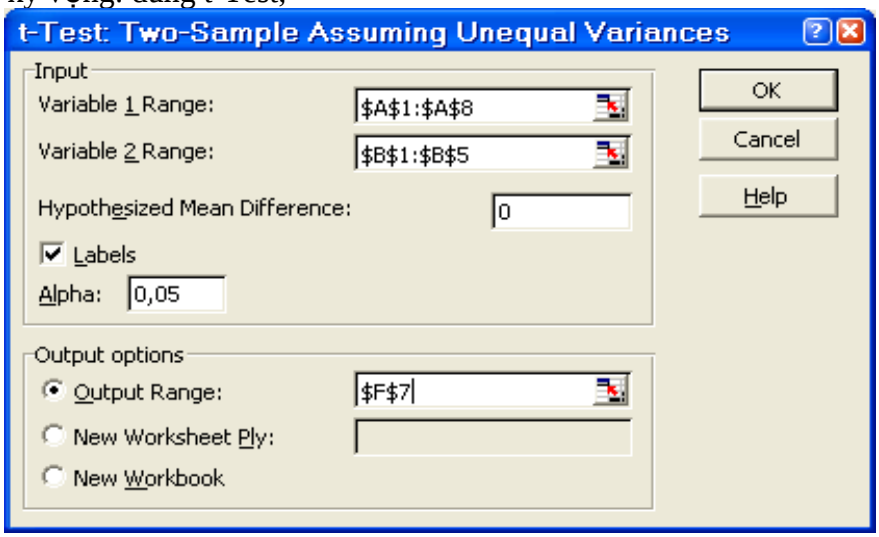


Hình 2

F-Test Two-Sample for Variances		
	Khu vực A	Khu vực B
Mean	36,08714	46,7625

Variance	16,65162	27,01269167
Observations (số mẫu quan sát)	7	4
df (bậc tự do = số mẫu -1)	6	3
F (Phân vị Fisher của kiểm định)	0,616437	
P(F<=f) one-tail	0,280739	
F Critical one-tail (Phân vị Fisher tới hạn)	0,102254	

Khi |F| <= |Fc| chấp nhận 2 phương sai có cùng độ chính xác.
Khi |F| > |Fc| 2 phương sai có độ chính xác khác nhau.
-So sánh 2 kỳ vọng: dùng t-Test,



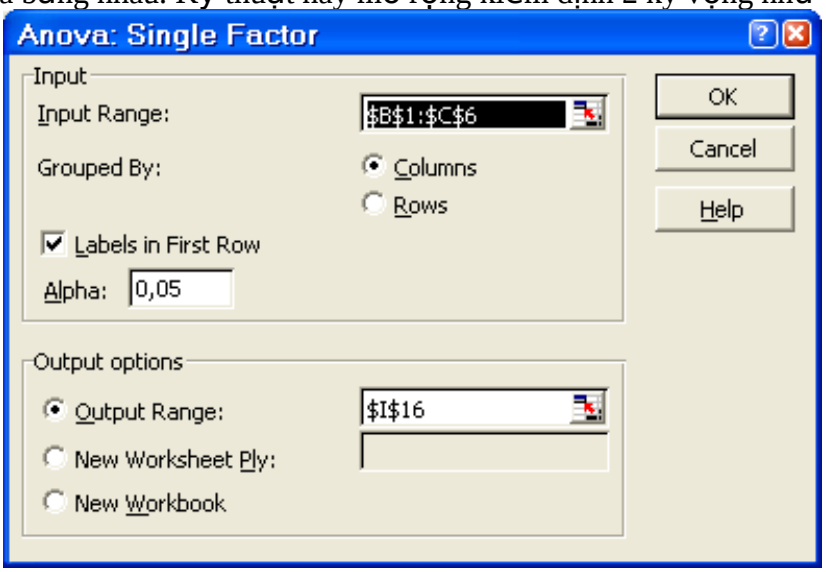
Hình 3

Hypothesized Mean Difference: giả định sai khác kỳ vọng = 0.
Có 3 kiểm định khác nhau dựa trên phương sai (có được do sử dụng F test).
* t-Test: two-sample assuming equal variances, dùng kiểm định khi phương sai cùng độ chính xác, kích thước các mẫu có thể khác nhau. Có thể dùng tìm hai mẫu có kỳ vọng bằng nhau.
* t-Test: two-sample assuming unequal variances, dùng kiểm định khi phương sai cùng độ chính xác, kích thước các mẫu có thể khác nhau.. Thường dùng trong nghiên cứu và thực nghiệm, có thể dùng kiểm định các mẫu trước và sau điều trị bệnh.

* t-Test: pair two sample for means: không giả thiết cùng phương sai, kích thước các mẫu phải bằng nhau nhau. Có thể dùng kiểm định các mẫu quan sát tự nhiên trước và sau khi thực nghiệm.
Với số liệu cho ở ví dụ trên, kết quả:

t-Test: Two-Sample Assuming Unequal Variances		
	Khu vực A	Khu vực B
Mean	36,087143	46,7625
Variance	16,651624	27,012692
Observations	7	4
Hypothesized Mean Difference	0	
Df	5	
t Stat	-3,532645	
P(T<=t) one-tail	0,0083463	
t Critical one-tail	2,0150492	
P(T<=t) two-tail	0,0166927	
t Critical two-tail	2,5705776	

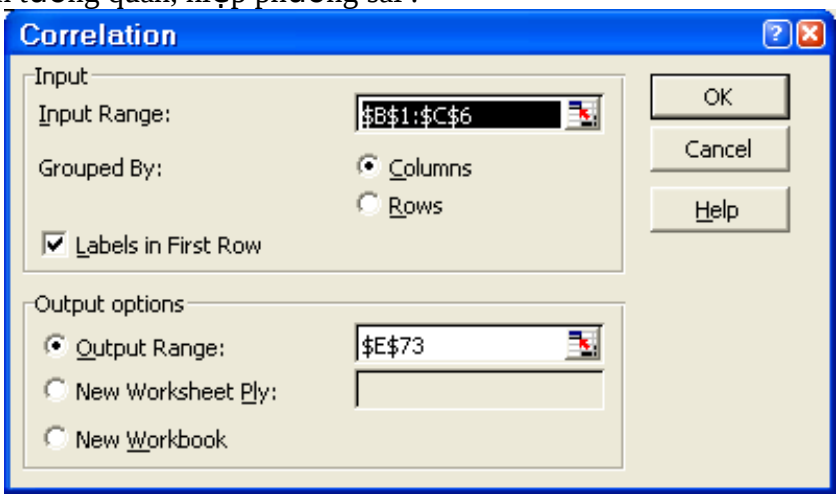
Trong đó:
t Stat - Phân vị Student của kiểm định.
t critical one tail: Phân vị Student tới hạn 1 phía(tra bảng với mức ý nghĩa a=5%).
t critical two tail: Phân vị Student tới hạn 2 phía(tra bảng với mức ý nghĩa a=2,5%).
Khi |t Stat| <= |t critical| chấp nhận giả thuyết 2 kỳ vọng bằng nhau.
Khi |t Stat| > |t critical |2 kỳ vọng khác nhau ở mức có ý nghĩa.
P(T<=t): mức ý nghĩa 1 và 2 phía.
Phân tích phương sai:
ANOVA: analysis of variance.
Có 3 loại phân tích tùy thuộc vào số các nhân tố và số các mẫu.
- Single Factor analysis: Kiểm định với giả thiết rằng kỳ vọng (trung bình) của 2 hoặc nhiều mẫu là bằng nhau. Kỹ thuật này mở rộng kiểm định 2 kỳ vọng như T-test.



Hình 4

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Trước khi phun	5	456	91,2	276,7		
Sau khi phun	5	465	93	185,5		
ANOVA						
Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	8,1	1	8,1	0,035049762	0,85615237	5,317644991
Within Groups	1848,8	8	231,1			
Total	1856,9	9				

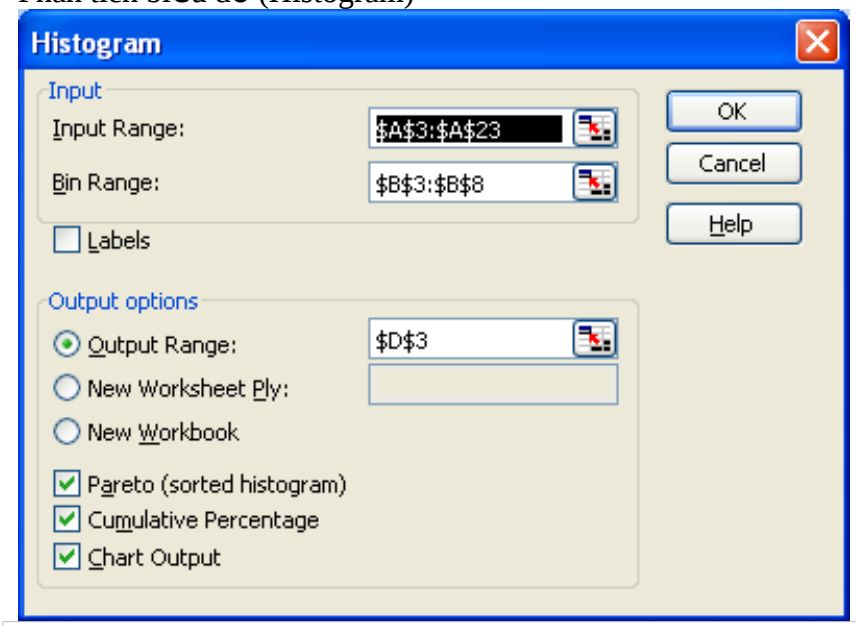
SS: Sum Square - Tổng bình phương độ lệch.
df: bậc tự do; dfG = k-1; dfW = n-k.
MS: Mean Square: Tổng bình phương độ lệch của kỳ vọng.
MSG = SSG/ dfG; MSW = SSW/ dfW.
F: Phân vị Fisher của kiểm định = MSG/MSW
P-value: Giá trị xác suất.
F crit: Phân vị Fisher tới hạn của dfG, dfW,
Khi F càng nhỏ thì P càng lớn và Mean càng gần bằng nhau.
|F|<|Fc|: chấp nhận giả thuyết mean các nhóm bằng nhau ở mức ý nghĩa
|F|>|Fc|: bác bỏ giả thuyết mean các nhóm bằng nhau ở mức ý nghĩa
việc xác định mean nào lớn hơn muốn chính xác cần tiến hành trong SPSS.
- Two-Factor With Replication (lặp lại): mở rộng của Single Factor gồm nhiều hơn cùng 1 mẫu cho mỗi nhóm dữ liệu.
- Two-Factor Without Replication: phân tích phương sai 2 nhân tố không bao gồm nhiều hơn cùng 1 mẫu cho mỗi nhóm, giả thiết kỳ vọng từ 2 hoặc nhiều mẫu là bằng nhau, là mở rộng của kiểm định 2 kỳ vọng như T- test.
Phân tích tương quan, hiệp phương sai :



Hình 5

	Trước khi phun	Sau khi phun
Trước khi phun	1	
Sau khi phun	0,984301907	1

Hệ số tương quan và hiệp phương sai, dùng đo mối liên hệ giữa 2 tập dữ liệu. Có thể dùng để xác định khả năng 2 miền dữ liệu chuyển đổi lẫn nhau, tương quan tuyệt đối (1), tương quan phủ định, hoặc không có mối liên hệ nào (0).
Dùng hàm: CORREL(Array1; array2).
COVAR(Array1; array2).
Tất cả các giá trị trên đều có thể tính trực tiếp từ các hàm thống kê có trong Excel, tuy vậy kết quả khoảng tin cậy có sự sai khác giữa tính toán theo hàm và theo phân tích.
Chú ý:
Khi phân tích, các số liệu cùng nhóm cần được xếp trên 1 hàng hoặc 1 cột, nếu không kết quả sẽ sai.
Khi tính toán, số liệu có thể xếp theo mảng (nhiều dòng và cột) công thức vẫn cho kết quả đúng.
Các kết quả có thể sai khác khi dùng các vertion Excel khác nhau
Các lựa chọn khác có thể chọn để tham khảo, thực hiện theo chỉ dẫn.
Phân tích biểu đồ (Histogram)



Số liệu		Nhóm							
41	69	50		Nhóm	Frequency	Cumulative %	Nhóm	Frequency	Cumulative %
47	72	60		50	3	13,64%	80	5	22,73%