

BÀI 1- NHẬT ĐỒ VÀ THỐNG KÊ

I- NỘI DUNG

Việc đầu tiên của xử lý dữ liệu là khảo sát và tính các thống kê cơ bản cho các biến. Trong sinh học có 2 loại biến : biến định tính và biến định lượng. Mỗi loại biến lại bao gồm nhiều loại như: biến định tính gồm biến nhị nguyên (Binary) chỉ lấy 2 giá trị (1 và 0 hay có và không), biến phân loại (hay định danh Nominal) gồm 1 số loại không sắp xếp thứ tự và biến thứ hạng (Ordinal). Biến định lượng gồm biến có thang đo khoảng cách (Interval scale) là biến trong đó hiệu số 2 giá trị có ý nghĩa còn tỷ số thì không có ý nghĩa, giá trị 0 chỉ mang tính quy ước, biến có thang đo tỷ số (Ratio scale) có gốc 0 có đơn vị đo và cả tỷ số lẫn hiệu số đều có ý nghĩa. Sau khi có dữ liệu thì cần làm một số việc:

a- SẮP XẾP SỐ LIỆU

Trường hợp ít số liệu thì để nguyên dãy số liệu hoặc sắp xếp lại theo thứ tự (Sort), trường hợp nhiều số liệu thì dùng 2 dãy: dãy các số liệu khác nhau và dãy số lần gặp (tần số). Trường hợp có rất nhiều số liệu thì chia khoảng rồi đếm số điểm rơi vào từng khoảng (tần số). Khi xử lý sẽ lấy điểm giữa làm đại diện cho khoảng.

Giá trị	Giá trị	Tần số	Khoảng cách	Điểm giữa	Tần số
x_1	x_1	m_1	$[x_0 - x_0 + h]$	x_1	m_1
x_2	x_2	m_2	$[x_0 + h - x_0 + 2h]$	x_2	m_2
...	...	...	...	...	...
x_n	x_k	m_k	$[x_0 + (k-1)h - x_0 + kh]$	x_k	m_k

b- ĐỒ THỊ

Có nhiều kiểu đồ thị để khảo sát sơ bộ dãy số liệu

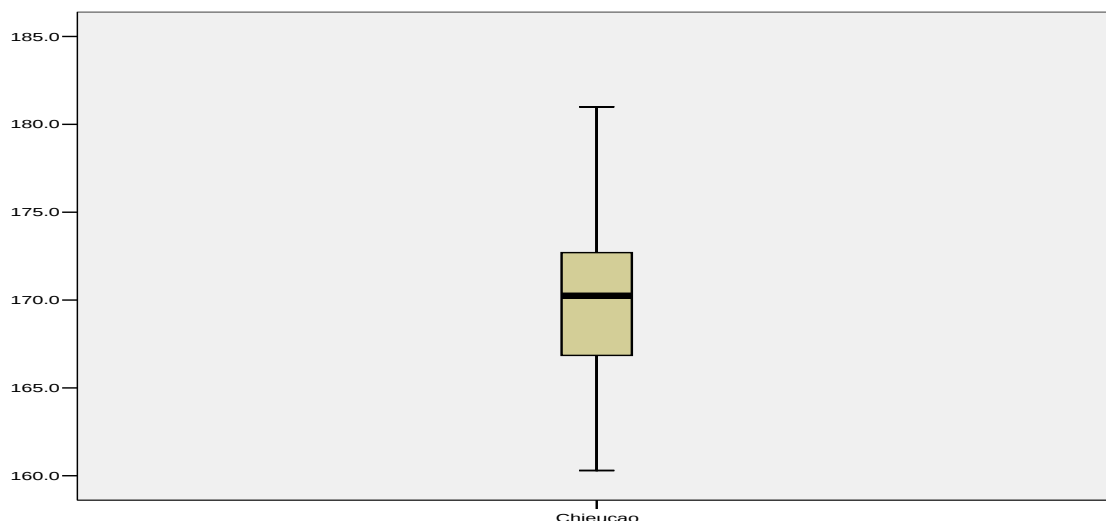
Kiểu cành và lá. Vẽ đồ thị thể hiện tần số của việc chia khoảng dãy số liệu (160-161), (162- 163), (164- 165), (166-167), (168-169) . . .).

Cành (Stem) viết tắt trị nguyên 160, 170, . . . còn lá (Leaf) là các số lẻ 0, 1, 2, . . .

Chieucao Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	16 .	0
5.00	16 .	22233
10.00	16 .	4444444555
22.00	16 .	6666666666777777777777
8.00	16 .	88888899
17.00	17 .	00000000111111111
18.00	17 .	222222222222233333
11.00	17 .	44444455555
5.00	17 .	66667
2.00	17 .	89
1.00	18 .	1
Stem width: 10		
Each leaf: 1 case(s)		

Kiểu hộp với đường trung vị Me ở giữa, đáy hộp là các đường tứ phân vị, hai ria kéo dài đến số to nhất và nhỏ nhất nếu các số này cách Me không quá 1,5 lần khoảng cách giữa các tứ phân vị (chiều dài hộp). Các điểm cách xa quá sẽ vẽ riêng từng điểm



Kiểu nhật đồ (Histogram) có thêm đường cong chuẩn.

Mục đích của nhật đồ: Chia lớp khi có rất nhiều số liệu để thấy rõ các nét đặc trưng cơ bản của dãy số liệu sau đó kiểm tra tính chuẩn

Cách làm: Gọi khoảng cách từ trị nhỏ nhất(Min) đến trị lớn nhất(Max) là khoảng biến động. Chia khoảng biến động thành một số khoảng nhỏ và đếm số số liệu nằm trong từng khoảng (tần số).

Về mặt đồ thể hiện tần số. Sơ bộ đánh giá số liệu có phân phối chuẩn hay không (các đỉnh của các chữ nhật thể hiện tần số có nằm gần đường cong mật độ chuẩn hay không). Các phần mềm thống kê lớn đều có các phần kiểm tra tính chuẩn theo các kiểm định Kolgomorov-Smirnov, Shapiro-Wilk, Ryan-Joiner, Anderson-Darling . . . Các kiểm định này thường so sánh hàm phân phối thực nghiệm và hàm phân phối chuẩn từ đó đưa ra một thống kê thể hiện sự sai khác kèm theo xác suất (P-value) để làm căn cứ có chấp nhận giả thiết dãy số liệu phân phối chuẩn hay không?

Nếu $P\text{-value} \leq \text{mức ý nghĩa } \alpha$ thì không chấp nhận tính chuẩn

Nếu $P\text{-value} > \text{mức ý nghĩa } \alpha$ thì chấp nhận tính chuẩn.

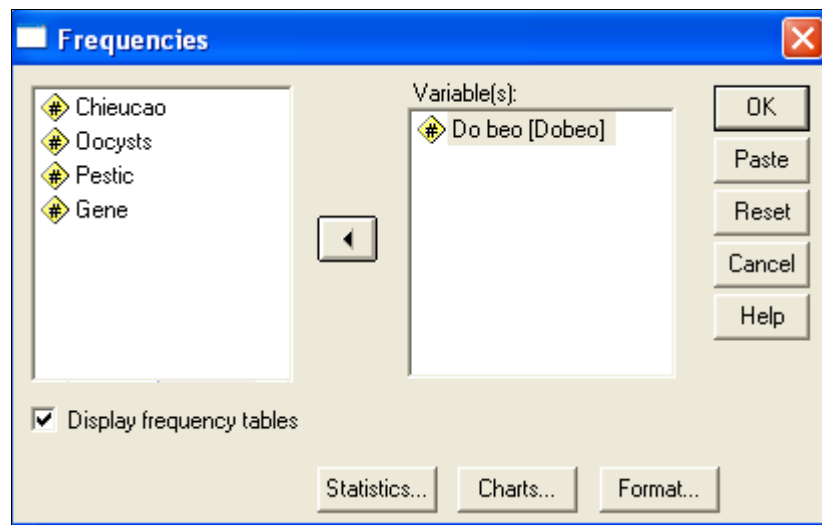
Tính chuẩn còn được kiểm tra qua đường cong mật độ chuẩn, qua đồ thị hàm phân phối thực nghiệm vẽ trên giấy xác suất (thường gọi là đường Q – Q quartiles – quartiles), v.v...

II- XỬ LÝ TRONG SPSS

Vào SPSS. Mở Worksheet Baitap1

Chọn menu Analyse Descriptive Statistics sau đó chọn một trong 3 cách Frequencies, Descriptives hay Explore

a- Frequencies chọn Dobeo đưa vào Variables sau đó vào Statistics (Thống kê) và Charts (Đồ thị), mỗi mục đều có các options để chọn lựa



Frequencies: Statistics

Percentile Values

☒ Quartiles

☐ Cut points for: 10 equal groups

☐ Percentile(s):

Add

Change

Remove

Central Tendency

☒ Mean

☒ Median

☒ Mode

☒ Sum

☐ Values are group midpoints

Dispersion

☒ Std. deviation

☒ Variance

☒ Range

☒ Minimum

☒ Maximum

☐ S.E. mean

Distribution

☒ Skewness

☒ Kurtosis

Continue

Cancel

Help

Frequencies: Charts

Chart Type

☐ None

☐ Bar charts

☐ Pie charts

☒ Histograms:

☒ With normal curve

Chart Values

☒ Frequencies

☐ Percentages

Continue

Cancel

Help

b- **Descriptives** Làm tương tự như frequencies

Descriptives

Variable(s):

Do beo [Dobeo]

Chieucao

Oocysts

Pestic

Gene

Save standardized values as variables

OK

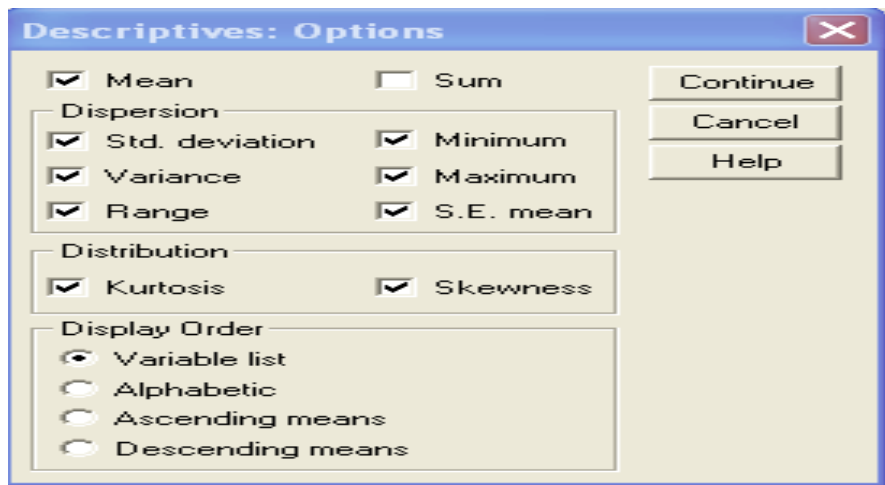
Paste

Reset

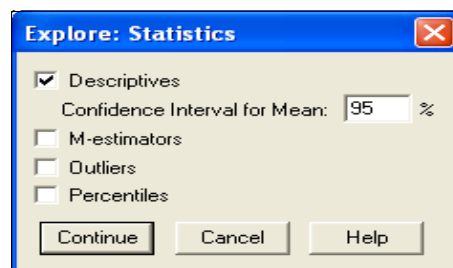
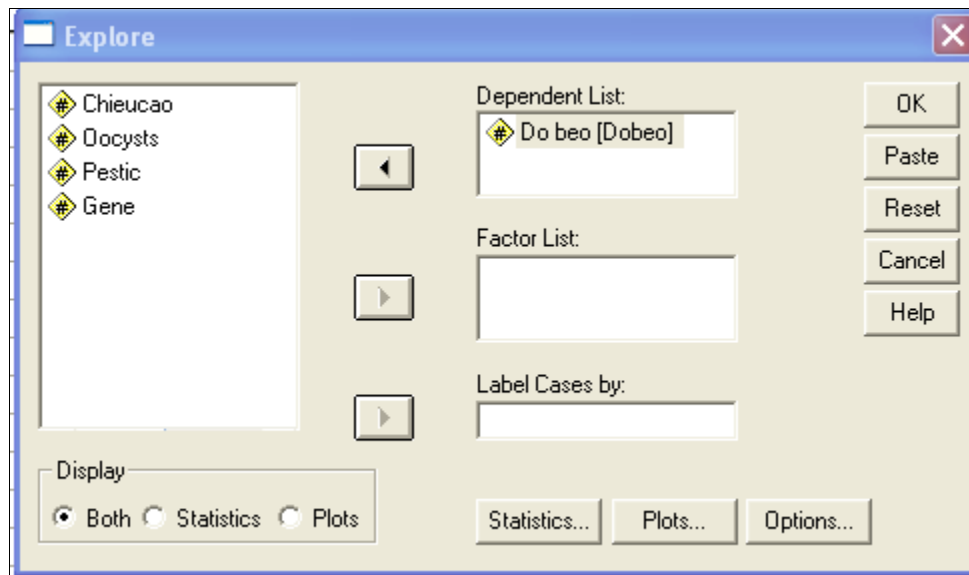
Cancel

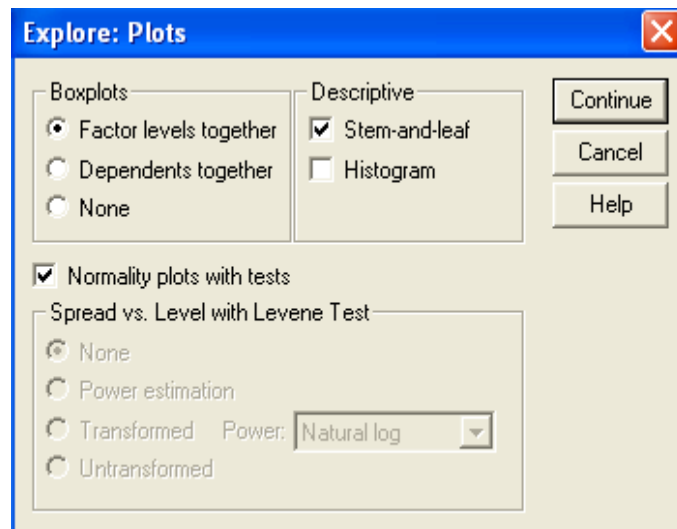
Help

Options...



- c- **Explore** Chọn Dobeo vào Dependent List, trong Statistics chọn Descriptives. Trong Plots chọn Histogram và Normality plots with Tests để vừa vẽ nhật đồ vừa kiểm định tính chuẩn





Các kết quả cơ bản

Các thống kê cơ bản: Tùy việc chọn Descriptives, frequencies hay explore và tùy theo options nên sẽ được toàn bộ hoặc một phần các thống kê cơ bản như Mean, Standard deviation, Median, Mode, Variance, . . .

Thí dụ trong Explore

			Statistic	Std. Error
Do béo	Mean		4.166	.0276
	95% Confidence Interval for Mean	Lower Bound	4.111	
		Upper Bound	4.221	
	5% Trimmed Mean		4.159	
	Median		4.145	
	Variance		.091	
	Std. Deviation		.3024	
	Minimum		3.5	
	Maximum		5.0	
	Range		1.5	
	Interquartile Range		.4	
	Skewness		.373	.221
	Kurtosis		.056	.438

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Do beo	.061	120	.200*	.987	120	.317

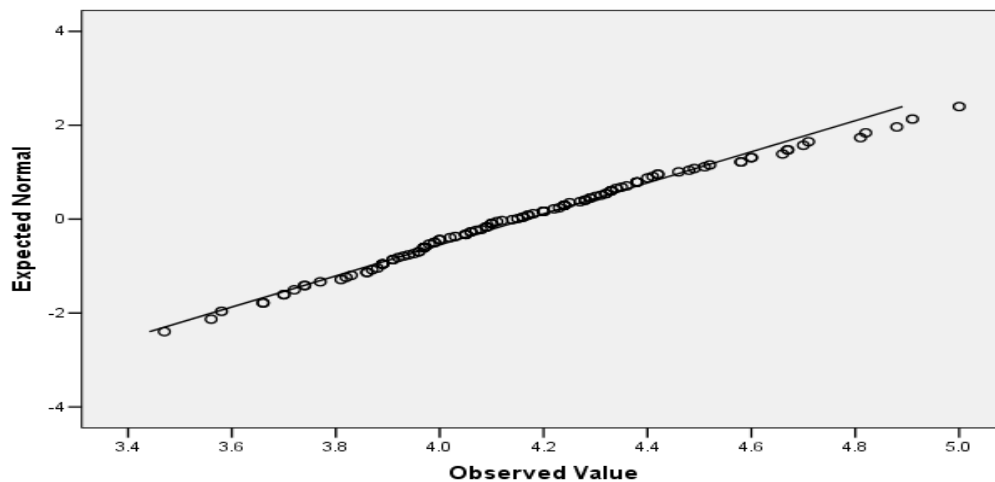
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Căn cứ vào mức Significant (0,200 và 0,317) của hai tiêu chuẩn kiểm định có thể kết luận: Chấp nhận tính chuẩn.

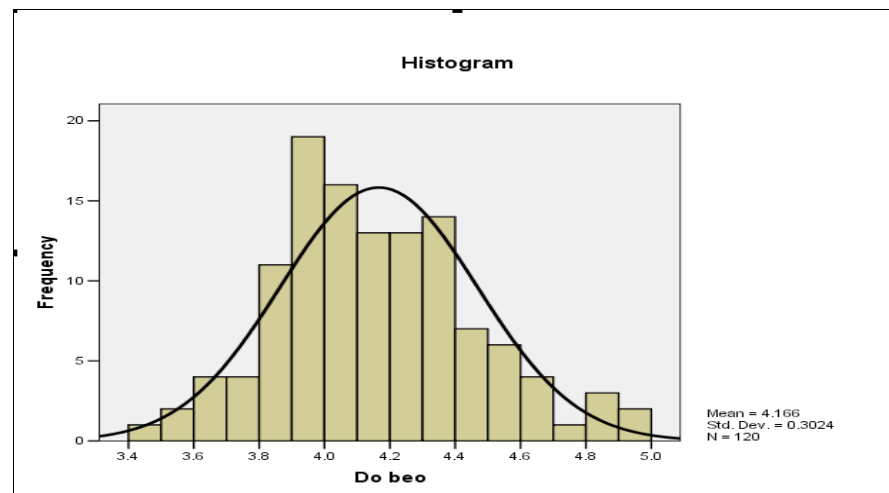
Đồ thị Q – Q để kiểm tra tính chuẩn

Normal Q-Q Plot of Do beo



Các điểm gần với đường thẳng nên có thể chấp nhận tính chuẩn.

Trong Frequencies có thể vẽ nhật đồ và đường cong chuẩn để đối chiếu



Để đưa biến không chuẩn về biến chuẩn có thể thực hiện các phép đổi biến, hay dùng nhất là phép bình phương, lấy căn bậc hai, lấy Lôgarit, biến đổi Arcsin.

Thí dụ biến oocysts không chuẩn

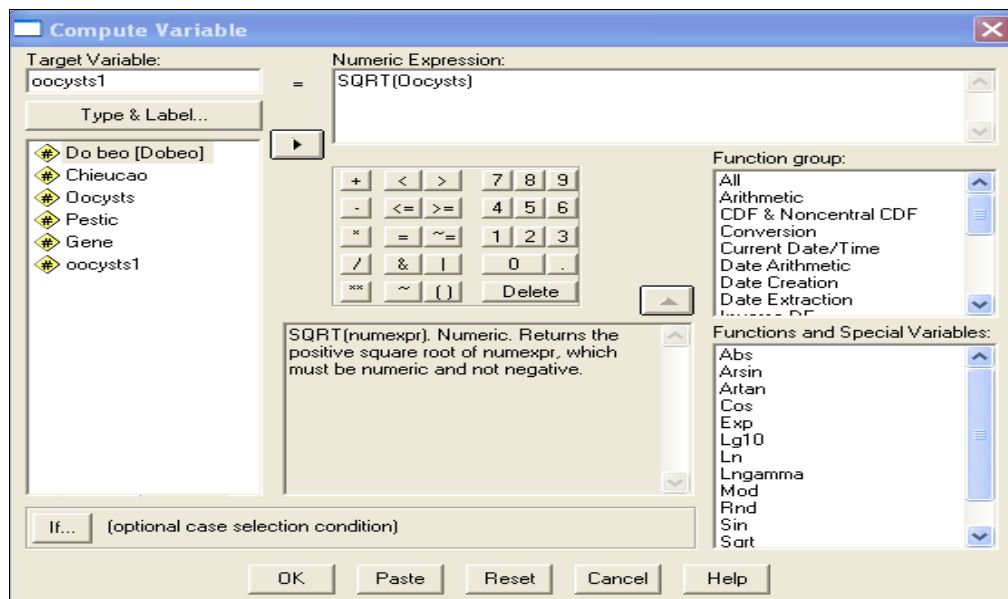
Tests of Normality ^a						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Oocysts	.140	100	.000	.862	100	.000

a. Lilliefors Significance Correction

Vào Transform Compute đặt tên biến mới oocysts1, chọn biểu thức thí dụ Sqrt (oocysts) sẽ được biến **oocysts1** (biến đổi căn bậc hai)

Vào Transform Compute đặt tên biến mới oocysts2, chọn biểu thức thí dụ oocysts *oocysts sẽ được biến **oocysts2** (Biến đổi bình phương)

Nếu chọn biến mới là oocysts3 sau đó chọn phép biến đổi log (oocysts) sẽ được phép biến **oocysts3**(Biến đổi logarit)



Biến oocysts1 và oocyst2 cũng không chuẩn. Thí dụ đối với oocysts1

Tests of Normality ^a						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
oocysts1	.083	100	.084	.952	100	.001

a. Lilliefors Significance Correction

Biến oocysts3 phân phối chuẩn

