

PHÂN PHỐI CHUẨN

Phân phối chuẩn (Normal distribution) được nêu ra bởi một người Anh gốc Pháp tên là Abraham de Moivre (1733). Sau đó Gauss, một nhà toán học người Đức, đã dùng luật phân phối chuẩn để nghiên cứu các dữ liệu về thiên văn học (1809) và do vậy cũng được gọi là phân phối Gauss. Theo từ điển bách khoa về khoa học thống kê, có lẽ người đầu tiên dùng từ “normal” là ông C.S Pierce (1780) vì vào thời đó người ta cho rằng mọi hiện tượng tự nhiên được coi như có phân phối chuẩn nhưng thật ra còn có những luật phân phối khác. Tuy vậy hầu hết lý thuyết thống kê được xây dựng trên nền tảng của phân phối chuẩn.

Như vậy từ “normal” được dùng theo thói quen nhưng thực ra không đúng, vì vậy trong tiếng Việt ta không thể dịch là phân phối “bình thường” mà gọi là phân phối chuẩn.

Hai thông số quan trọng trong một phân phối là giá trị trung tâm hay gọi là trung bình μ và phương sai σ^2 (hoặc độ lệch chuẩn σ) và thường biểu thị bằng $X \sim N(\mu, \sigma^2)$ (N viết tắt từ normal).

Nếu phân phối chuẩn được chuẩn hóa với trung bình $\mu=0$ và độ lệch chuẩn $\sigma=1$, được viết tắt là: $Z \sim N(\mu=0, \sigma=1)$, được gọi là phân phối chuẩn chuẩn hóa (standardized normal distribution) nghe có vẻ không được xuôi tai như tiếng Anh vì chữ normal được dịch là chuẩn rồi, do vậy dùng từ phân phối chuẩn tắc có vẻ ổn hơn!

Nói chung các đặc tính sinh trắc học của người khỏe mạnh (cân nặng, chiều cao, trị số mạch, huyết áp, đường máu, số lượng hồng cầu), thường tuân theo luật phân phối chuẩn. Ví dụ: xét nghiệm đường máu 100 người lớn khỏe mạnh các kết quả thu thập trong bảng 10.1.

Bảng 10.1 Kết quả đường máu (mg%) 100 người lớn khỏe mạnh

97	100	94	106	103	108	97	92	113	112
88	108	95	101	124	95	119	99	84	93
82	114	88	85	79	90	104	104	109	98
94	89	102	98	93	102	102	102	110	109
94	114	106	109	103	90	93	83	104	106
100	111	101	88	80	91	103	91	91	119
97	116	118	117	95	92	123	81	102	95
106	106	95	103	96	89	94	122	110	104
84	108	104	98	98	97	105	109	98	86
105	97	87	111	107	115	96	94	79	107

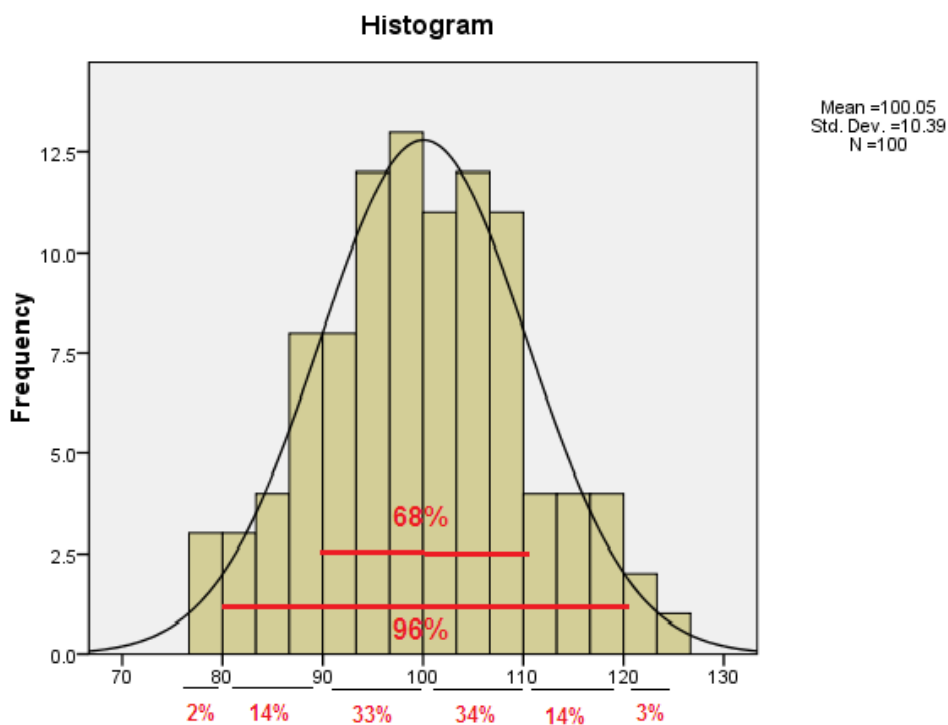
Bảng 10.2 Biểu đồ cuống-lá của đường máu:

Frequency	Stem & Leaf
2.00	7 . 99
6.00	8 . 012344
8.00	8 . 56788899
15.00	9 . 001112233344444
18.00	9 . 555556677777888889
18.00	10 . 001122222333344444
16.00	10 . 5566666778889999
8.00	11 . 00112344
6.00	11 . 567899
3.00	12 . 234

Nhìn vào biểu đồ cuống-lá ta thấy có:

- 2 người có trị đường máu <80mg%: 2%
- 14 người có trị đường máu 80-89mg%: 14%
- 33 người có trị đường máu 90-99mg%:33%
- 34 người có trị đường máu 100-109mg%: 34%
- 14 người có trị đường máu 110-119mg%: 13%
- 3 người có trị đường máu >120 mg%: 3%

Và biểu đồ tần suất (histogram) của phân phối đường máu (Biểu đồ 10.1):



Biểu đồ 10.1 Phân phối đường máu của 100 người lớn khỏe mạnh

Như vậy ta thấy phân phối lượng đường máu tuân theo luật chuẩn với trị số trung bình $\mu=100$ và độ lệch chuẩn $\sigma=10$ với:

68% giá trị quan sát nằm trong khoảng σ của μ .

95% giá trị quan sát nằm trong khoảng 2σ của μ .

99,7% giá trị quan sát nằm trong khoảng 3σ của μ .

(còn gọi là **luật 68-95-99,7**)

10.1 Hàm mật độ phân phối chuẩn

Hàm mật độ phân phối chuẩn (Normal density probability function) có dạng tổng quát như sau:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

Trong đó: $\pi = 3,14159\dots$
 $e = 2,71828\dots$ (cơ số logarit Neper)
 μ : trị số trung bình
 σ : độ lệch chuẩn

Biến ngẫu nhiên X có đơn vị là mg% bây giờ ta muốn chuyển đơn vị đo lường của biến số X theo đơn vị đo lường tổng quát cho mọi phân phối chuẩn nghĩa là theo đơn vị độ lệch chuẩn. Lúc đó phân phối chuẩn theo X sẽ trở thành phân phối chuẩn tắc (standadized normal distribution) với biến số mới là Z .

Muốn đổi hàm $y=f(x)$ ra hàm chuẩn tắc $y=f(z)$ ta đặt:

$$z = \frac{x - \mu}{\sigma}$$

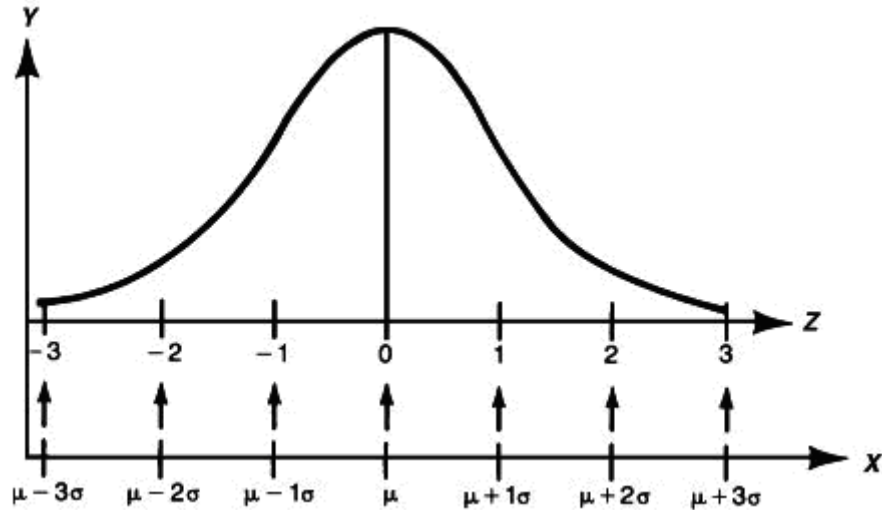
Thế $\mu=100$ và $\sigma=10$ ta có:

$$z = \frac{x - 100}{10}$$

Như vậy khi:

- $x=80 \rightarrow z=-2$
- $x=90 \rightarrow z=-1$
- $x=100 \rightarrow z=0$
- $x=110 \rightarrow z=+1$
- $x=120 \rightarrow z=+2$

Và đường cong chuẩn $y=f(z)$ sẽ là:



Sơ đồ 10.1 Biến đổi phân phối chuẩn X (trị trung bình μ , độ lệch chuẩn σ) thành phân phối chuẩn tắc Z (trị trung bình=0, độ lệch chuẩn=1)

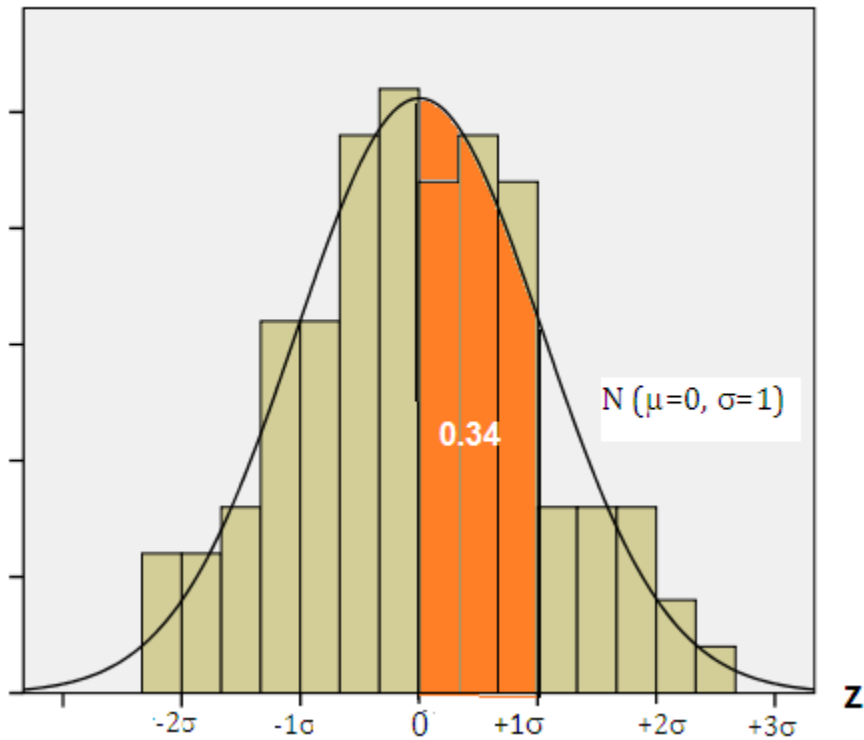
Tóm lại: Biến X tuân theo luật chuẩn với trung bình μ và phương sai σ^2 thường được viết tắt là: $X \sim N(\mu, \sigma^2)$ và biến Z tuân theo phân phối chuẩn tắc có $\mu=0$ và $\sigma^2=1$ được viết là $Z \sim N(0,1)$. Như vậy lúc này Z có đơn vị là độ lệch (ví dụ: 1, 2 hoặc 3 độ lệch chuẩn so với trị trung bình) và không tùy thuộc vào đơn vị đo lường theo biến X (ví dụ. mg% đường máu).

Phương trình đường cong chuẩn tắc theo Z sẽ là:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Lúc này muốn biết xác suất đường máu từ 100-110mg% (theo X) chỉ cần tính xác suất từ 0 đến 1 đơn vị độ lệch chuẩn theo Z hoặc tìm diện tích dưới đường cong từ 0 đến 1 (phần màu đậm-hình 3). Tích phân của hàm $f(z)$ từ $0 \rightarrow 1$ chính là diện tích dưới đường cong này. Trong thống kê gọi $F(Z)$ là hàm xác suất chuẩn tích lũy (cumulative normal probability function)

$f(z)$



Biểu đồ 10.2 Diện tích dưới đường cong chuẩn từ $0 \rightarrow +1\sigma$

$$f(z) = \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Công thức tính tích phân hàm $F(z)$ khá rắc rối thường ta dùng bảng Z-score (phần phụ lục) để tính. Xem bảng khi $z=0 \rightarrow z=1$: $F(z)=0,34$

$$P(0 \leq z \leq 1) = \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0,34$$

Như vậy xác suất $P(0 \leq Z \leq 1)$ là 0,34 hoặc xác suất những người có trị đường máu từ $X=100\text{mg}\%$ (tương đương với $Z=0$) đến $X=110 \text{mg}\%$ (tương đương với $Z=1$) là 34% (biểu đồ)

Các khoảng đặc biệt có thể tính nhằm xác suất:

$$P(-\infty \leq z \leq 0) = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0,5$$

$$P(0 \leq z \leq +\infty) = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0,5$$

$$P(-\infty \leq z \leq +\infty) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1$$

Một ví dụ khác: Muốn tính xác suất của z từ $-\infty$ đến 1,2 ta lấy: xác suất khoảng từ $-\infty$ đến 0 là $p=0,50$ cộng với xác suất khoảng từ 0 đến 1,2 là 0,38 (xem bảng z-score phần phụ lục), tổng cộng 2 xác suất này là 0,88 (tương đương 88% người có đường máu 115mg%) (1 đơn vị z bằng 10mg%)

10.2 Nhận biết một phân phối chuẩn trong SPSS

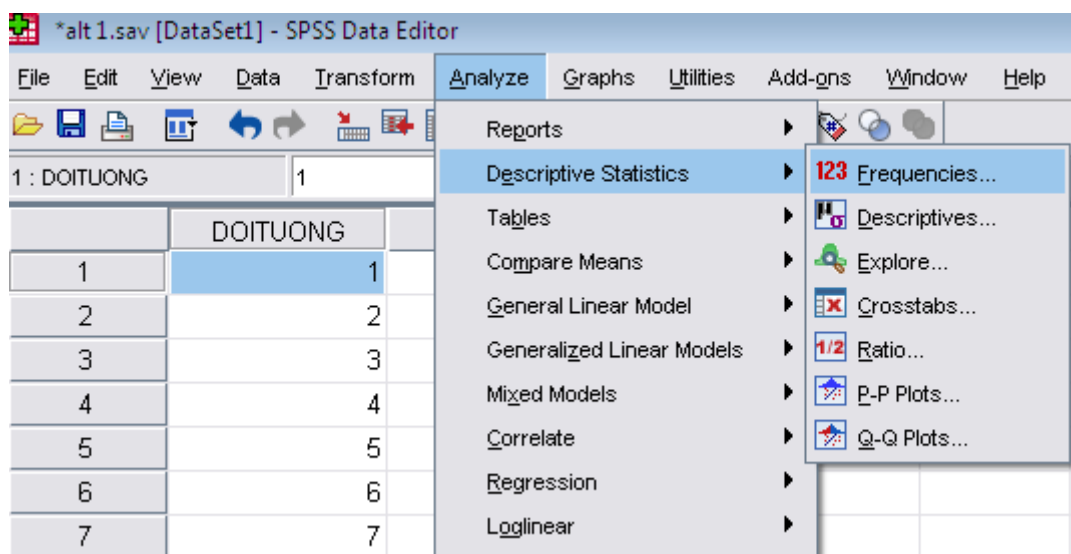
Có nhiều cách để đánh giá một phân phối chuẩn trong SPSS.

- (1) Đơn giản nhất là xem biểu đồ với đường cong chuẩn (Histograms with normal curve) với dạng hình chuông đối xứng với tần số cao nhất nằm ngay giữa và các tần số thấp dần nằm ở 2 bên. Trị trung bình (mean) và trung vị (median) gần bằng nhau và độ xệch (skewness) gần bằng zero.
- (2) Vẽ biểu đồ xác suất chuẩn (normal Q-Q plot). Phân phối chuẩn khi biểu đồ xác suất này có quan hệ tuyến tính (đường thẳng)
- (3) Dùng phép kiểm định Kolmogorov-Smirnov khi cỡ mẫu lớn hơn 50 hoặc phép kiểm Shapiro-Wilk khi cỡ mẫu nhỏ hơn 50. Được coi là có phân phối chuẩn khi mức ý nghĩa (Sig.) lớn hơn 0,05.

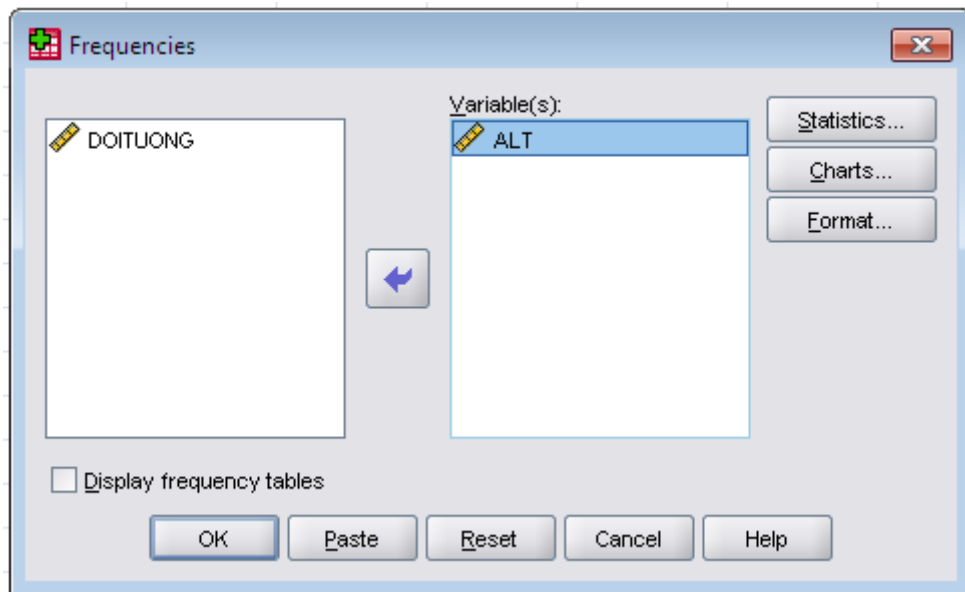
Ví dụ 1. Khảo sát men ALT (U/L) trên 30 người bình thường, kết quả được nhập vào SPSS như sau:

	DOITUONG	ALT
1	1	12
2	2	13
3	3	14
4	4	15
5	5	16
6	6	17
7	7	18
8	8	19
9	9	20
10	10	21
11	11	22
12	12	23
13	13	24
14	14	25
15	15	26
16	16	27
17	17	28
18	18	29
19	19	30
20	20	31
21	21	32
22	22	33
23	23	34
24	24	26
25	25	22
26	26	23
27	27	24
28	28	25
29	29	26
30	30	44

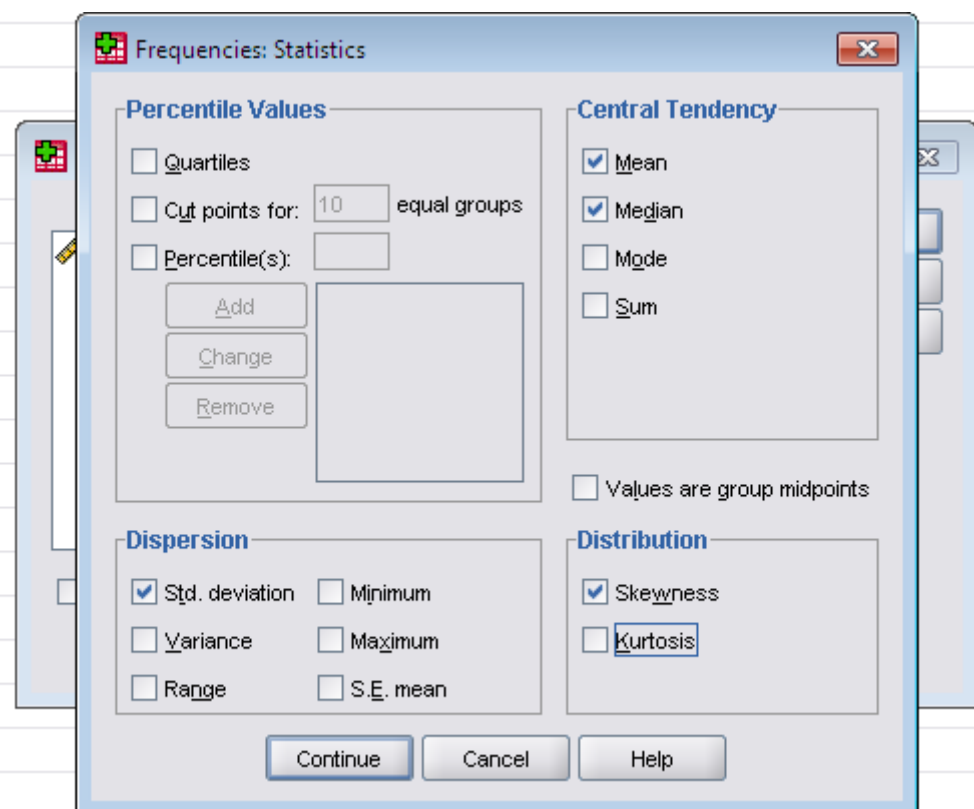
Và thực đơn Analyze > Descriptive Statistics > Frequencies...



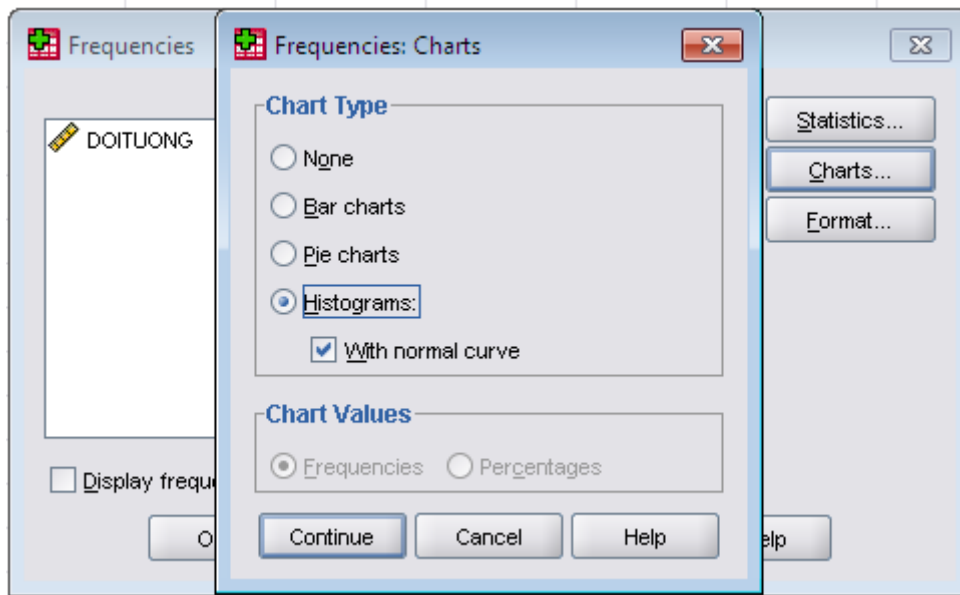
Mở màn hình Frequencies. Nhấp chuyển biến ALT từ ô bên trái vào ô Variable(s).



Nhấp hộp thoại Statistics... Vào màn hình Statistics, đánh dấu nháy vào 4 ô: Mean, Median, Std. deviation và Skewness và nhấp Continue



Nhấp tiếp hộp thoại Charts.. Đánh dấu vào ô tròn Histograms: và đánh dấu nháy vào ô With Normal curve, nhấp Continue. Nhấn OK sẽ cho kết quả sau.



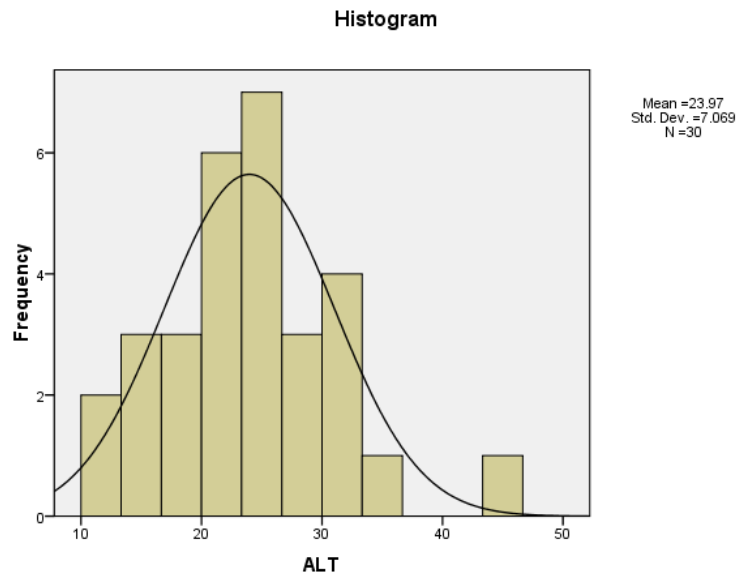
➔ **Frequencies**

[DataSet1] F:\BAIVIET\BAITAP\alt 1.sav

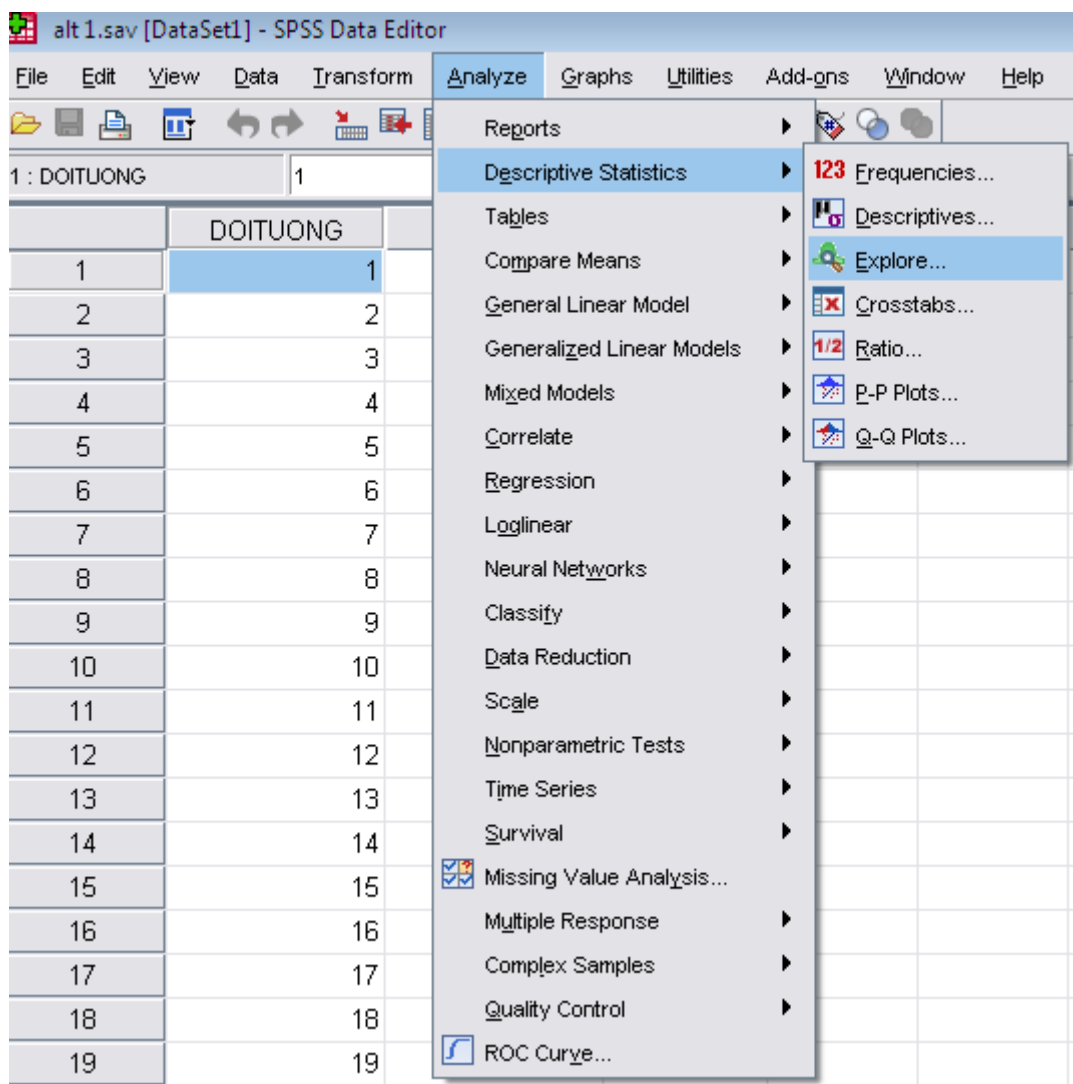
Statistics

ALT		
N	Valid	30
	Missing	0
Mean		23.97
Median		24.00
Std. Deviation		7.069
Skewness		.533
Std. Error of Skewness		.427
Kurtosis		.856
Std. Error of Kurtosis		.833

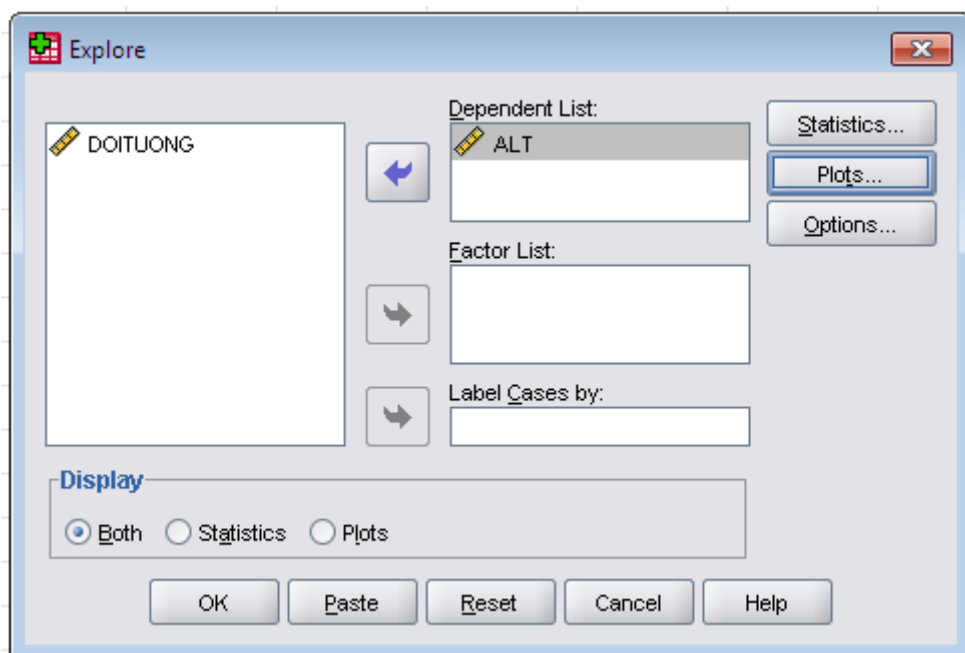
Trị trung bình (mean)= 23,97, trung vị (median)=24,00 và độ xiên(skewness)=0,533
 Trong phân phối này, trị số trung bình và trung vị gần bằng nhau và độ xiên dao động từ -1 đến +1. Như vậy đây có thể là một phân phối chuẩn. Thật vậy xem biểu đồ phân phối với đường cong chuẩn có dạng hình chuông, có trị trung bình là 23,97 và số liệu phân phối khá đều 2 bên.



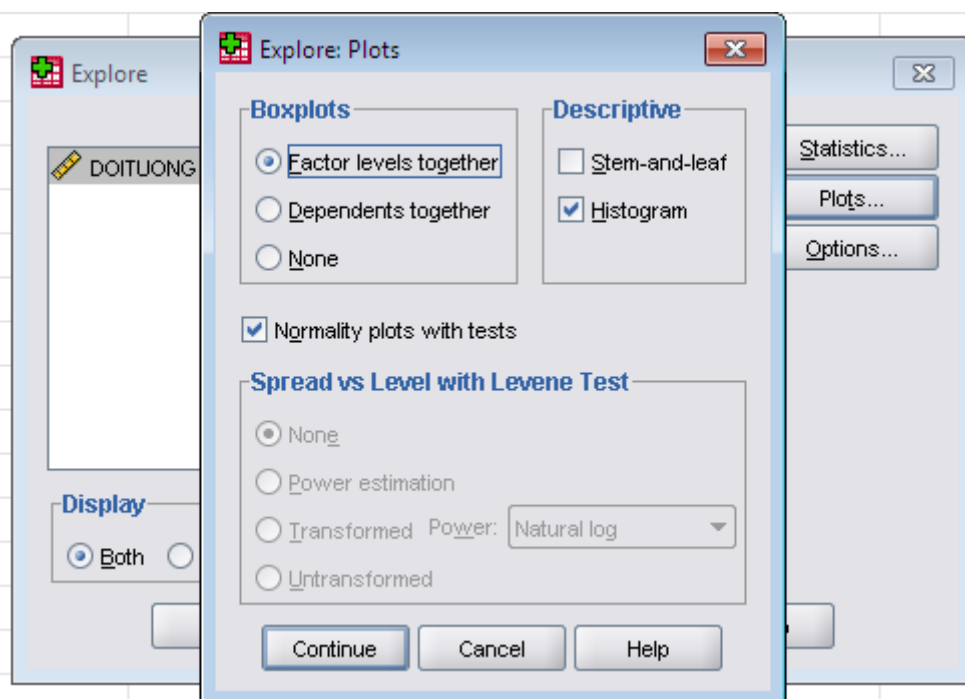
Để kiểm định Kolmogorov-Smirnov hoặc vẽ biểu đồ xác suất chuẩn Q-Q (Normal Q-Q plot) vào thực đơn: **Analyze > Descriptive Statistics > Explore...**



Khi xuất hiện màn hình Explore, chuyển ALT từ ô bên trái vào ô Dependent List:



Nhấn vào hộp thoại Plots. Sau khi màn hình Explore:Plots xuất hiện, nhấp dấu nháy vào ô Histogram và ô Normality plots with tests. Nhấp Continue và nhấp OK



Kết quả kiểm định phân phối chuẩn như sau:

Descriptives

			Statistic	Std. Error
ALT	Mean		23.97	1.291
	95% Confidence Interval for Mean	Lower Bound	21.33	
		Upper Bound	26.61	
	5% Trimmed Mean		23.69	
	Median		24.00	
	Variance		49.964	
	Std. Deviation		7.069	
	Minimum		12	
	Maximum		44	
	Range		32	
	Interquartile Range		10	
	Skewness		.533	.427
	Kurtosis		.856	.833

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ALT	.087	30	.200 [*]	.971	30	.571

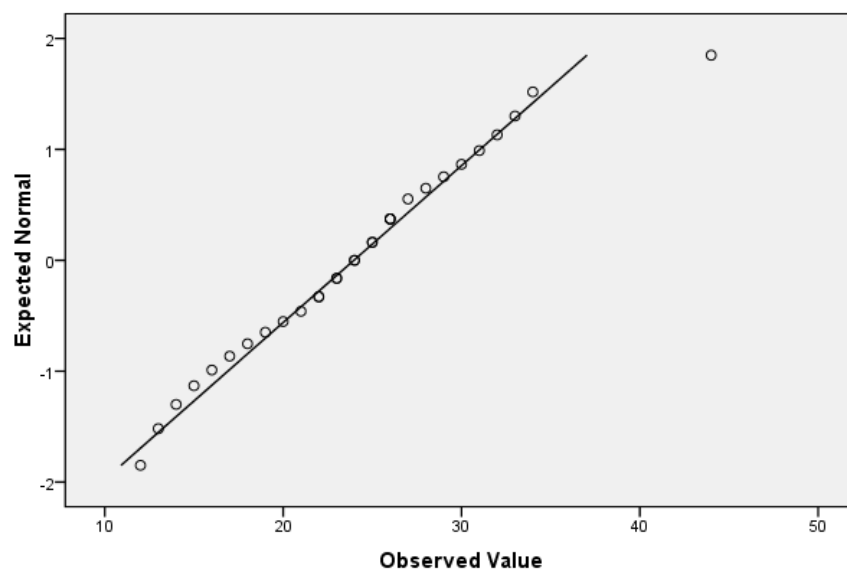
a. Lilliefors Significance Correction

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

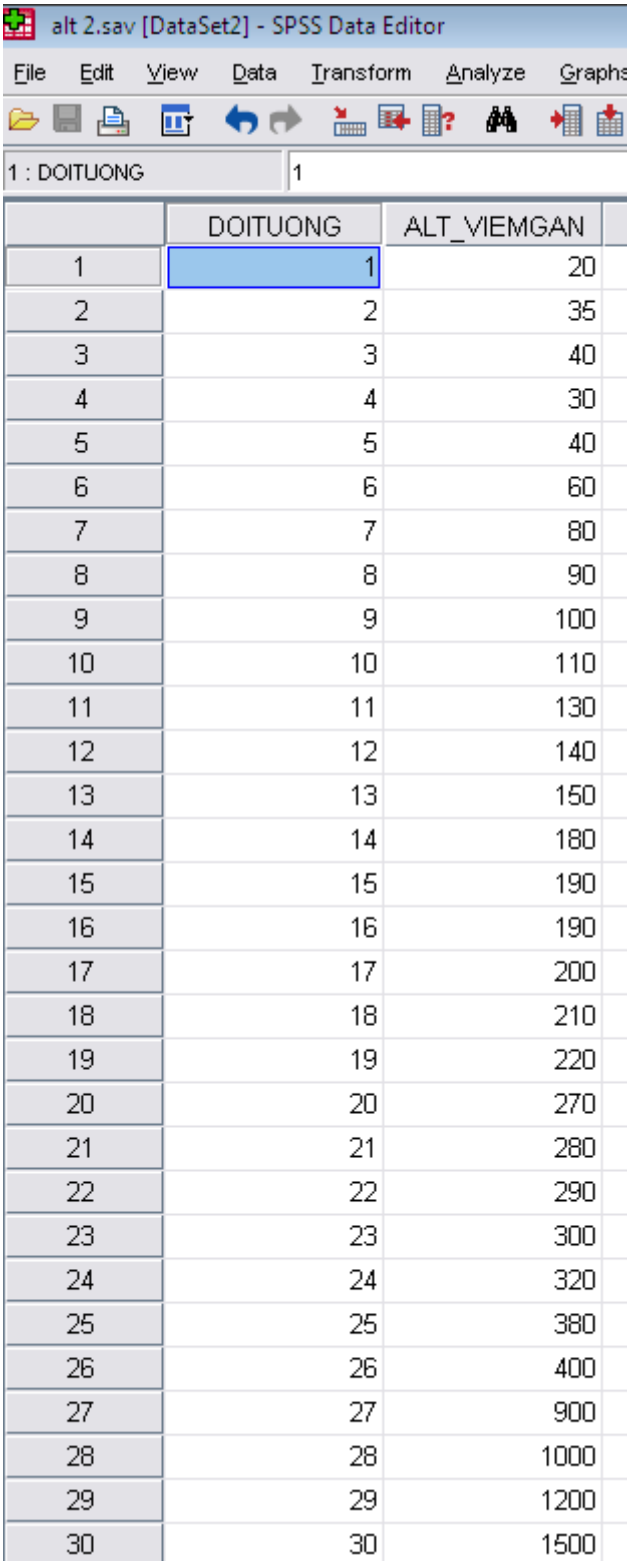
Vì cỡ mẫu 30 (nhỏ hơn 50), dùng kiểm định Shapiro-Wilk với Sig.=0,571 (lớn hơn 0,05). Chứng tỏ phân phối này là phân phối chuẩn.

Xem biểu đồ Normal Q-Q Plot bên dưới, các trị số quan sát và trị số mong đợi đều nằm gần trên đường thẳng

Normal Q-Q Plot of ALT



Ví dụ 2. Khảo sát men ALT (U/L) trên 30 người mắc viêm gan siêu vi B mãn tính



The image shows a screenshot of the SPSS Data Editor window. The title bar reads "alt 2.sav [DataSet2] - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, and Graphs. Below the menu bar is a toolbar with various icons. The main window displays a data grid with 30 rows and 2 columns. The first column is labeled "DOITUONG" and the second column is labeled "ALT_VIEMGAN". The data points are as follows:

	DOITUONG	ALT_VIEMGAN
1	1	20
2	2	35
3	3	40
4	4	30
5	5	40
6	6	60
7	7	80
8	8	90
9	9	100
10	10	110
11	11	130
12	12	140
13	13	150
14	14	180
15	15	190
16	16	190
17	17	200
18	18	210
19	19	220
20	20	270
21	21	280
22	22	290
23	23	300
24	24	320
25	25	380
26	26	400
27	27	900
28	28	1000
29	29	1200
30	30	1500

Vào Analyze > Descriptive > Frequencies như phần trên
Kết quả

→ Frequencies

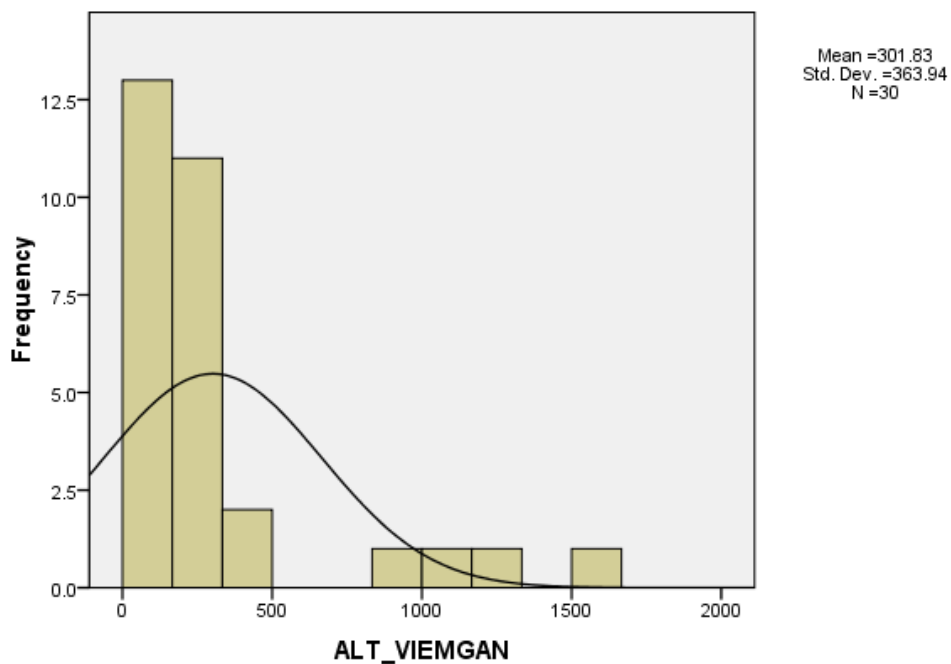
[DataSet2] F:\BAIVIET\BAITAP\alt 2.sav

Statistics

ALT_VIEMGAN		
N	Valid	30
	Missing	0
Mean		301.83
Median		190.00
Std. Deviation		363.940
Skewness		2.172
Std. Error of Skewness		.427

Phân phối này có trung bình (301,82) và trung vị (190,00) khá khác biệt. Hơn nữa độ xệch (2,17) lớn hơn +1, vì vậy có thể không phải là phân phối chuẩn. Thật vậy, xem biểu đồ với đường cong chuẩn cho thấy số liệu không phân phối đều, tập trung nhiều về phía đuôi trái và ít về phía đuôi phải (xệch phải)

Histogram

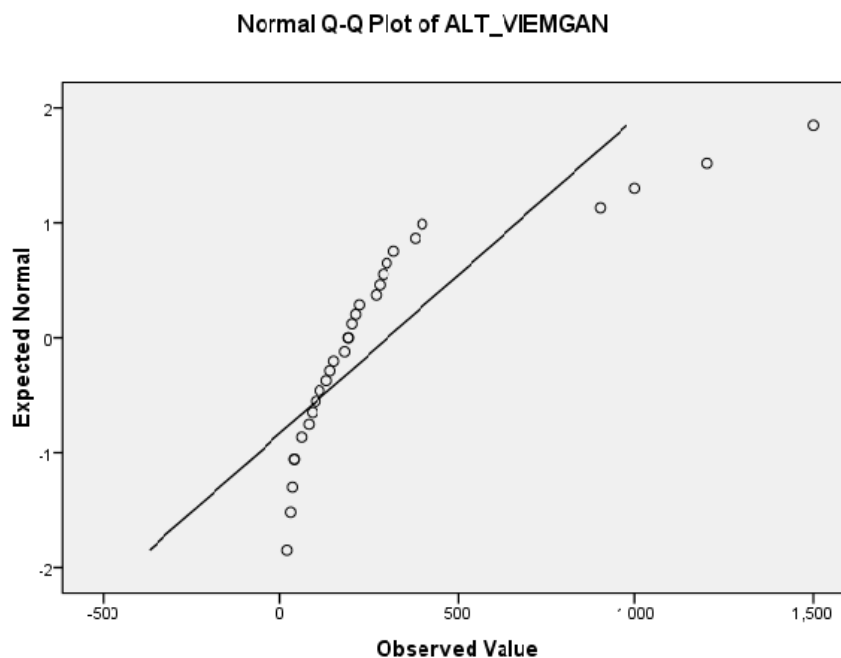


ĐỂ rõ hơn vào Analyze>Descriptives> Explore để xem kiểm định Kolmogorov-Smirnov và biểu đồ Normal Q-Q plot

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ALT_VIEMGAN	.280	30	.000	.690	30	.000

a. Lilliefors Significance Correction

Kiểm định Shapiro-Wilk với Sig=0,000 (nhỏ hơn 0,05 → bác bỏ giả thuyết không). Như vậy phân phối này không phải là phân phối chuẩn. Trên biểu đồ Q-Q plot ta thấy sự liên hệ giữa trị số quan sát và trị số mong đợi không nằm trên đường thẳng.



Tài liệu tham khảo:

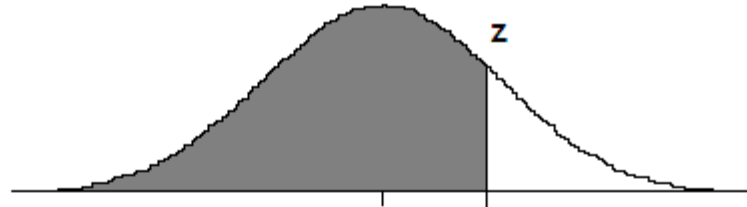
1. Armitage P. and Berry G. The normal distribution, in Statistical methods in medical research, 3rd edition, Backwell Scientific publication 1994, pp;66-71.
2. Altman DG. The normal distribution. statistic notes. BMJ 1995; 310:298.
3. Website: <http://www.stat.wvu.edu/SRS/Modules/Normal/normal.html> truy cập ngày 12/02/09.

Phụ lục 10.1 Tính xác suất p (hàng) theo Z (cột)

Ví dụ: $Z=0 \rightarrow p= 0.50$

$Z=1 \rightarrow p= 0.84$

$Z=2 \rightarrow p= 0.97$



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990