14. Статистическа обработка на данни

Извадка наричаме частта от генералната съвкупност, която се изследва. Броят на елементите в извадката се нарича обем на извадката.

Репрезентативна (представителна) извадка е тази, която отразява достоверно генералната съвкупност. Ако изборът на елементите става по случаен начин, то извадката се нарича случайна.

От опит е установено, че **репрезентативни са случайните извадки с голям обем.**

Затова ще разглеждаме само случайни извадки.

Тъй като изборът на елементи от генералната съвкупност е случаен, то променливата която се изучава, ще разглеждаме като случайна величина X с

неизвестен закон, а резултатите от всяко измерване (или броене) – възможни стойности x_1 , x_2 ,..., x_n на случайни величини X_1 , X_2 ,…, X_n със същия закон на разпределение.

Така при изследване на количествената променлива X на генералната съвкупност:

1. Величината X се счита за случайна величина с неизвестен закон;
2. извадката се представя като многомерна случайна величина (X_1,…,X_n) , чийто компоненти X_i={резултат от i-тото наблюдение} имат разпределение, съвпадащо с разпределението на X;
3. наблюдаваните стойности x_1, x_2 ,..., x_n са реализация на величините X_1, X_2 ,…, X_n, т.е. наблюдавани са събитията $X_1 = x_1$, $X_2 = x_2$ и т.н.

Следователно, теоретична основа на статистиката е теорията на вероятностите.

Вариационен ред

Наблюдаваните стойности, наредени по големина: $\hat{x}_1 \le \hat{x}_2 \le \ldots \le \hat{x}_n$.

Задача 1. Трудовия стж на 10 работника във фирма е 7, 14, 11, 17, 21, 15, 18, 20, 30, 15.
Вариационният ред на данните е: 7, 11, 14, 15, 15, 17, 18, 20, 21, 30.

Честотна таблица

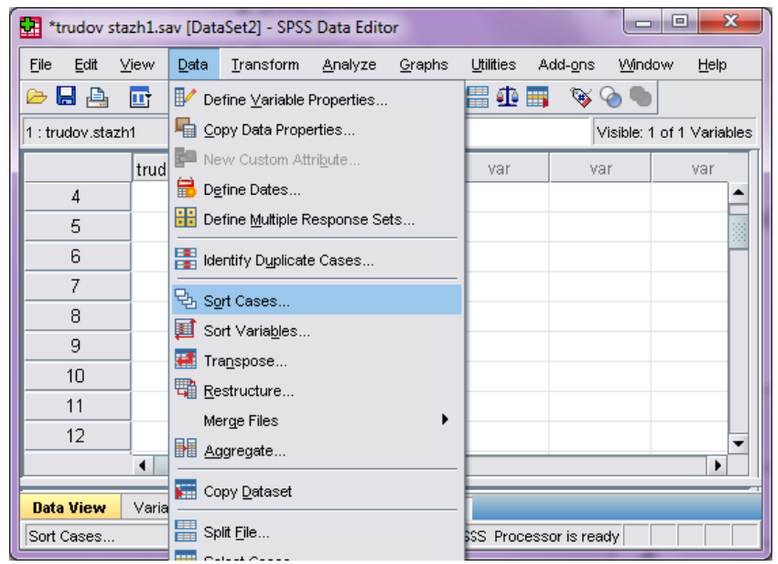
Таблица от вида $\begin{array}{c|cccc} X & x_1 & x_2 & \ldots & x_k \\ \hline m_i & m_1 & m_2 & \ldots & m_k \end{array}$,

m_i, $\nu_i = \dfrac{m_i}{n}$ - честота и относителна честота на x_i, $\sum_{i=1}^{k} m_i = n$ - обем на извадката;

$\gamma_i = \sum_{j=1}^{n} \nu_j$ - кумулативна (натрупана) честота на x_i.

Задача 2. Трудовия стаж на 22 работника във фирма е 12, 14, 12, 11,14, 17, 11, 7, 15, 15, 14, 17, 11, 7, 12, 14, 12, 14, 11, 7, 14, 12.

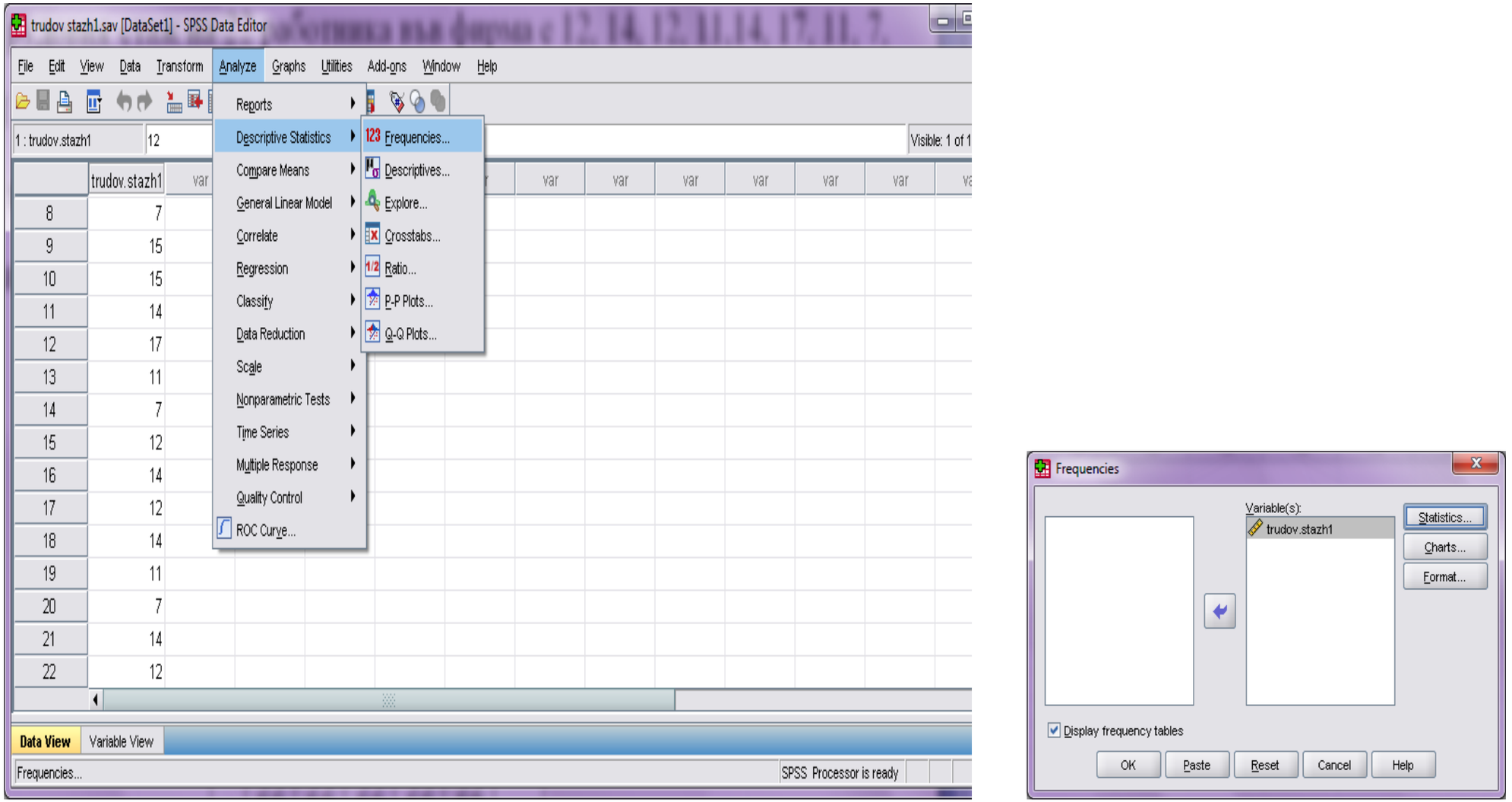
За вариационен ред в SPSS: Data->Sort Cases



Фиг. 1.

Честотна таблица, таблица на относителните честоти и на натрупаните (кумулативни) относителни честоти.

x_i	**7**	**11**	**12**	**14**	**15**	**17**
n_i	3	4	5	6	2	2
ν_i	$\frac{3}{22}=0,136$	$\frac{4}{22}=0,182$	$\frac{5}{22}=0,227$	$\frac{3}{11}=0,273$	$\frac{1}{11}=0,091$	$\frac{1}{11}=0,091$
γ_i	$\frac{3}{22}=0,136$	$\frac{7}{22}=0,318$	$\frac{12}{22}=0,545$	$\frac{18}{22}=0,818$	$\frac{20}{22}=0,909$	1


trudov stazh1.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Compare Means
General Linear Model
Correlate
Regression
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Multiple Response
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
1 : trudov.stazh1
12
Visible: 1 of 1
trudov.stazh1
var
8 7
9 15
10 15
11 14
12 17
13 11
14 7
15 12
16 14
17 12
18 14
19 11
20 7
21 14
22 12
Data View
Variable View
Frequencies...
SPSS Processor is ready
Frequencies
Variable(s):
trudov.stazh1
Statistics...
Charts...
Format...
Display frequency tables
OK
Paste
Reset
Cancel
Help

Фиг. 2.

trudov.stazh1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7	3	13,6	13,6	13,6
	11	4	18,2	18,2	31,8
	12	5	22,7	22,7	54,5
	14	6	27,3	27,3	81,8
	15	2	9,1	9,1	90,9
	17	2	9,1	9,1	100,0
	Total	22	100,0	100,0	

Фиг. 3.

Групиран статистически ред.

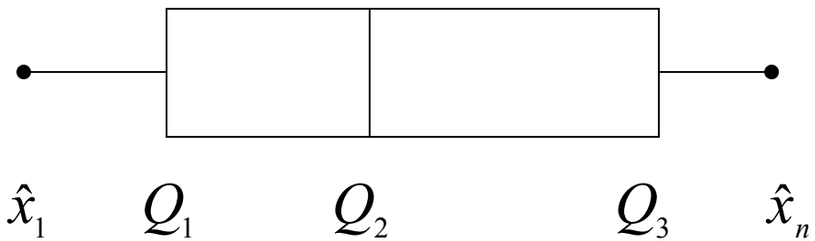
Използва се, когато данните са много. Тогава вариационният ред се разделя на интервали, обикновено с равни дължини, като във всеки интервал трябва да има данни.

Броят на подинтервалите k може да се намери от неравенството min k: $2^k > n$. Определят се средите на получените интервали – m_i. Построява се таблица на групирания статистически ред, като на първия ред се записват стойностите на m_i, а на втория – честотите, т.е. броят на данните, които се намират в дадения подинтервал.

M_i	m_1	m_2	m_3	...	m_k
f_i	f_1	f_2	f_3	...	f_k

Квантил от ред p $(0 < p < 1)$ – число x_p, за което $100p\%$ от извадката са по-малки от него.

Бокс-плот – диаграма от вида, където



Фиг. 4.

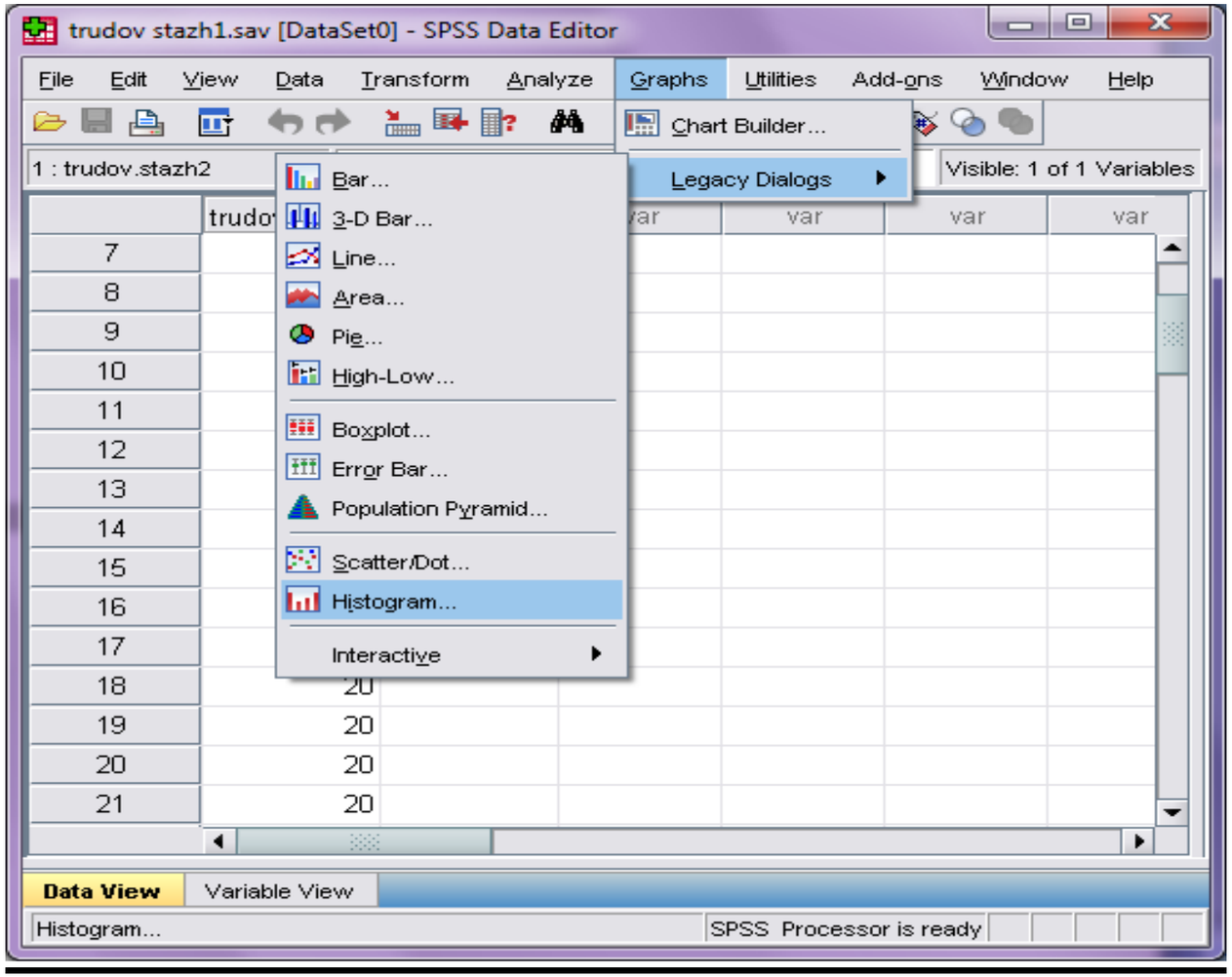
$$Q_1 = x_{0,25},\ Q_2 = x_{0,5} = Md,\ Q_3 = x_{0,75}$$

Задача 3. Трудовия стаж на 60 работника във фирма е 7, 14, 11, 17, 17, 21, 15, 15, 18, 11, 15, 16, 20, 14, 25, 25, 35, 27, 24, 24, 31, 28, 33, 25, 28, 33, 24, 7, 13, 21, 16, 22, 23, 28, 23, 27, 27, 31, 34, 34, 37, 41, 20, 41, 13, 20, 20, 25, 30, 22, 22, 34, 31, 34, 24, 29, 34, 36, 8, 37.

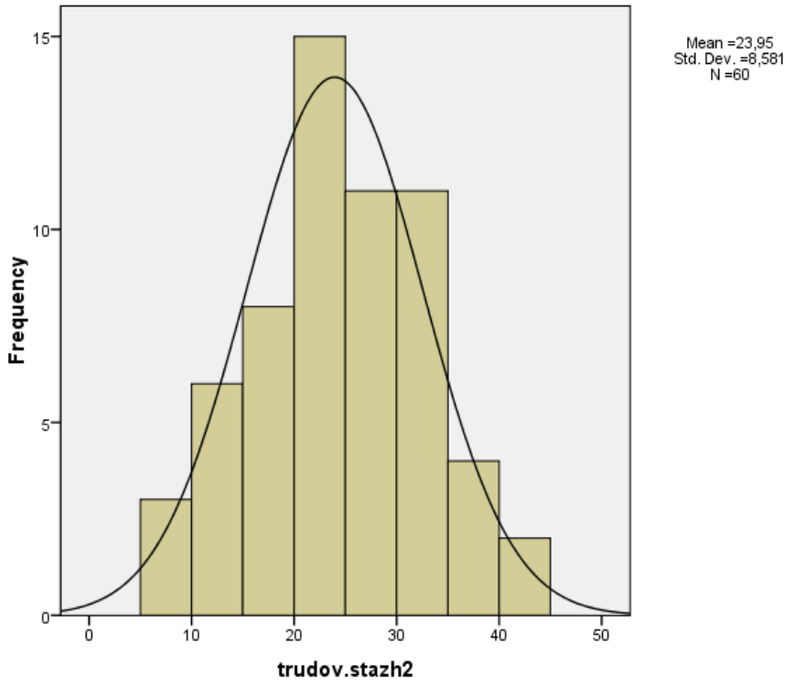
Брой данни n = 60. Приблизително $2^6 > 60$, затова избираме брой подинтервали k=6. Дължините на интервалите нека са равни: $h=(x_{max}-x_{min})/k=(41 - 7)/6 = 5, 666 \approx 6$, т.е. h = 6. Групираният статистически ред е даден в долната таблица.

интервал	(6,12]	(12, 18]	(18, 24]	(24, 30]	(30, 36]	(36, 42]
m_i	9	15	21	27	33	39
f_i	5	11	12	15	12	5
v_i	$\frac{5}{60}$	$\frac{11}{60}$	$\frac{12}{60}$	$\frac{15}{60}$	$\frac{12}{60}$	$\frac{5}{60}$

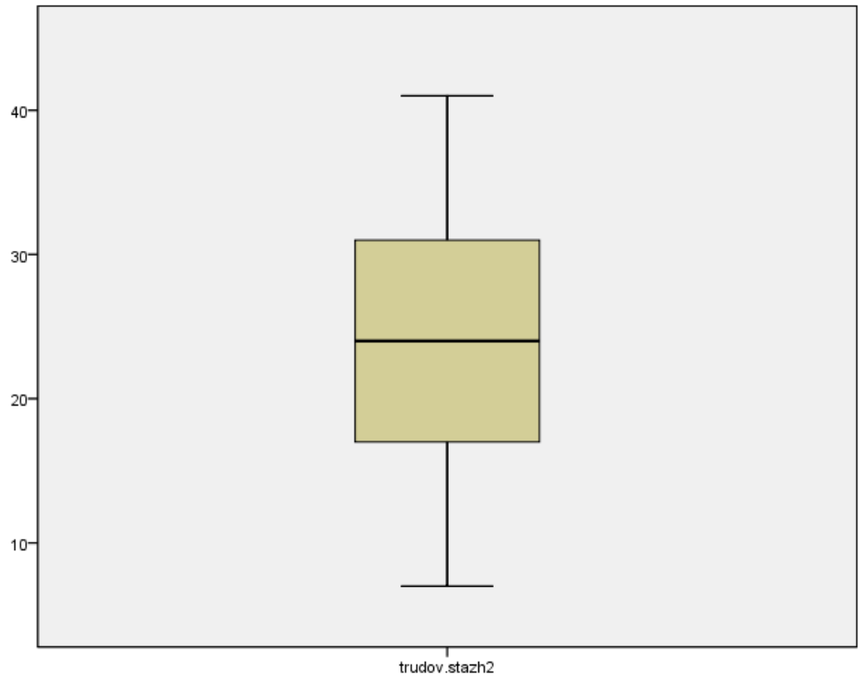
Статистическият ред се използва за построяване на полигон и хистограма на данните.


trudov stazh1.sav [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Chart Builder...
Legacy Dialogs
1 : trudov.stazh2
Visible: 1 of 1 Variables
Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...
Interactive
var
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
20
Data View
Variable View
Histogram...
SPSS Processor is ready

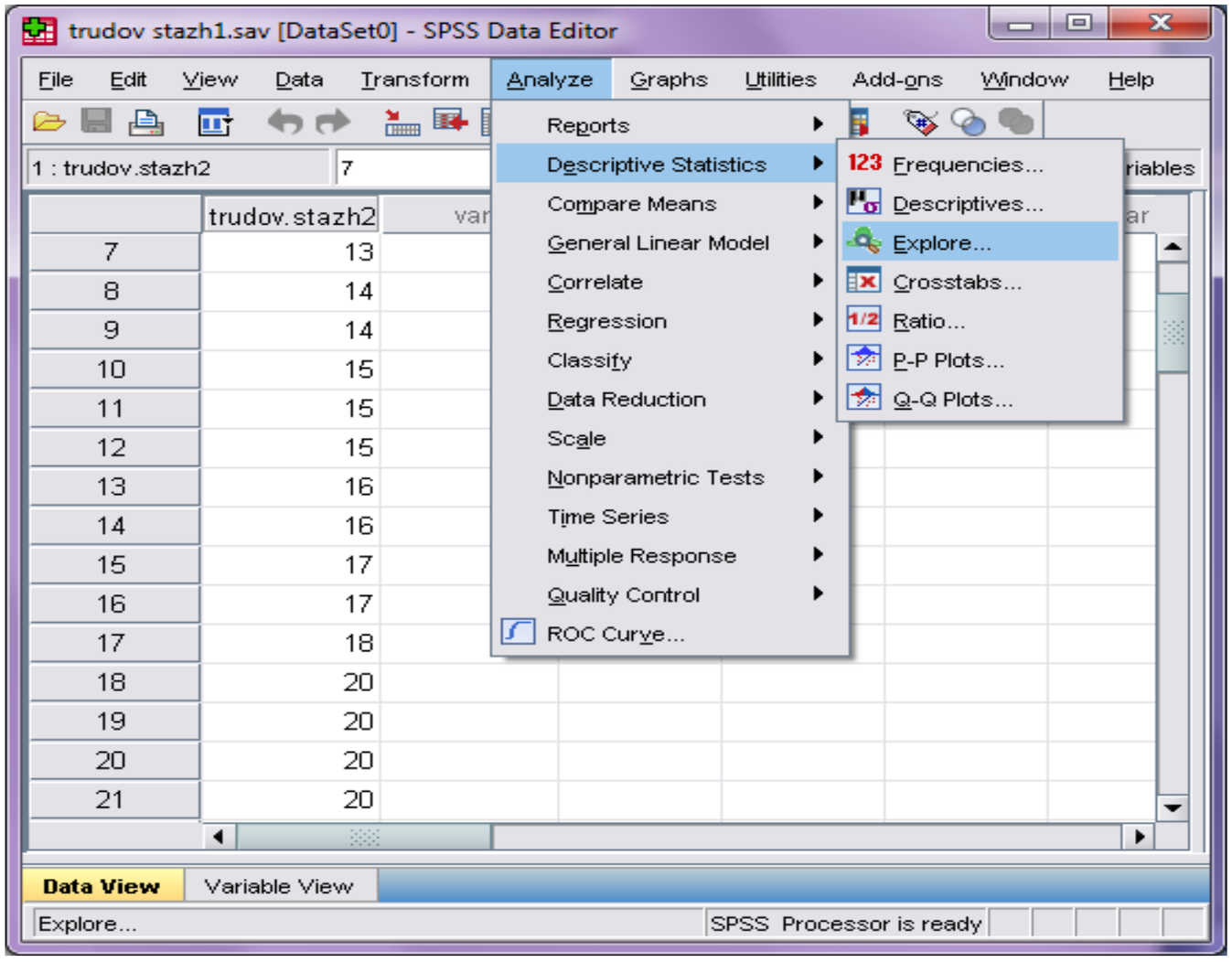
Фиг. 5.


Mean =23,95
Std. Dev. =8,581
N =60
Frequency
15
10
5
0
0
10
20
30
40
50
trudov.stazh2

Фиг. 6.


40
30
20
10
trudov.stazh2

Фиг. 7.


trudov stazh1.sav [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : trudov.stazh2
7
trudov.stazh2
7 13
8 14
9 14
10 15
11 15
12 15
13 16
14 16
15 17
16 17
17 18
18 20
19 20
20 20
21 20
Reports
Descriptive Statistics
Compare Means
General Linear Model
Correlate
Regression
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Multiple Response
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
Data View
Variable View
Explore...
SPSS Processor is ready

Фиг. 8.

Точкови оценки (Числени характеристики на извадка)

$\bar{x}$ – **извадково средно**. Тя е ефективна оценка на математическото очакване на генералната съвкупност (ГС):

- за вариационен ред $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$
- за статистически ред: $\bar{x} = \frac{x_1 n_1 + x_2 n_2 + ... + x_k n_k}{n}$
- за групиран статистически ред: $\bar{x} = \frac{m_1 f_1 + m_2 f_2 + ... + m_k f_k}{n}$

Md - медиана на извадката. Тя е оценка на медианата на ГС. Нейната стойност се явява среда на извадката, при което броят на елементите по-малки или равни на Md

е равен на броя на елементите по-големи или равни на Md. Това е 0,5 квантил на извадката.

Mo - мода на извадката. Тя е най-често срещаната стойност в извадката.

- от статистическия ред: **Mo** $= x_i$, за което честотата n_i е най-голяма.

s^2 - дисперсия на извадката (извадкова дисперсия). Тя е ефективна оценка на дисперсията на ГС:

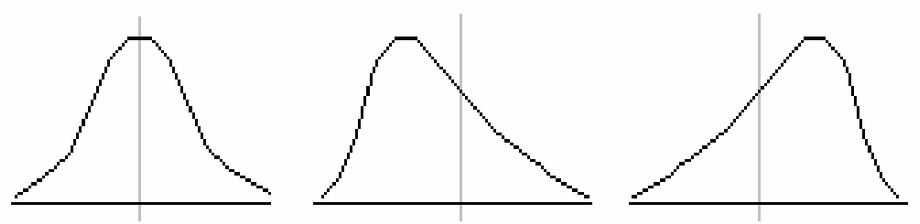
- за вариационен ред: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- за статистически ред: $s^2 = \frac{1}{n_1 + n_2 + \ldots + n_k - 1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2$

- за групиран статистически ред: $s^2 = \frac{1}{f_1 + f_2 + \ldots + f_k - 1}\sum_{i=1}^{k} f_i (m_i - \bar{x})^2$

s – стандартното отклонение на извадката е ефективна оценка на стандартното отклонение на ГС. $s = \sqrt{s^2}$.

извадков коефициент на асиметрия – дава представа за симетричността на разпределението.



Sk = 0 симетрично

Sk>0 дясно асиметрично

Sk<0 ляво асиметрично

- за вариационен ред: $Sk = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$

- за статистически ред: $Sk = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^3}{s^3}$

- за групиран статистически ред: $Sk = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^3}{s^3}$

извадков коефициент на ексцес – дава представа за островърхостта на разпределението и за тежестта на опашките

Ku = 0 Ku > 0 Ku <0

Нормалното разпределение е симетрично, т.е. коефициента на асиметрия на нормалното разпределение е 0. Ексцеса на нормалното разпределение също е 0.

Задача 4. Да се намерят числените характеристики на данните от ***задача 2***.

x_i	**7**	**11**	**12**	**14**	**15**	**17**
n_i	3	4	5	6	2	2

$$\overline{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n} = \frac{7.3 + 11.4 + 12.5 + 14.6 + 15.2 + 17.2}{3 + 4 + 5 + 6 + 2 + 2} = 12{,}41$$

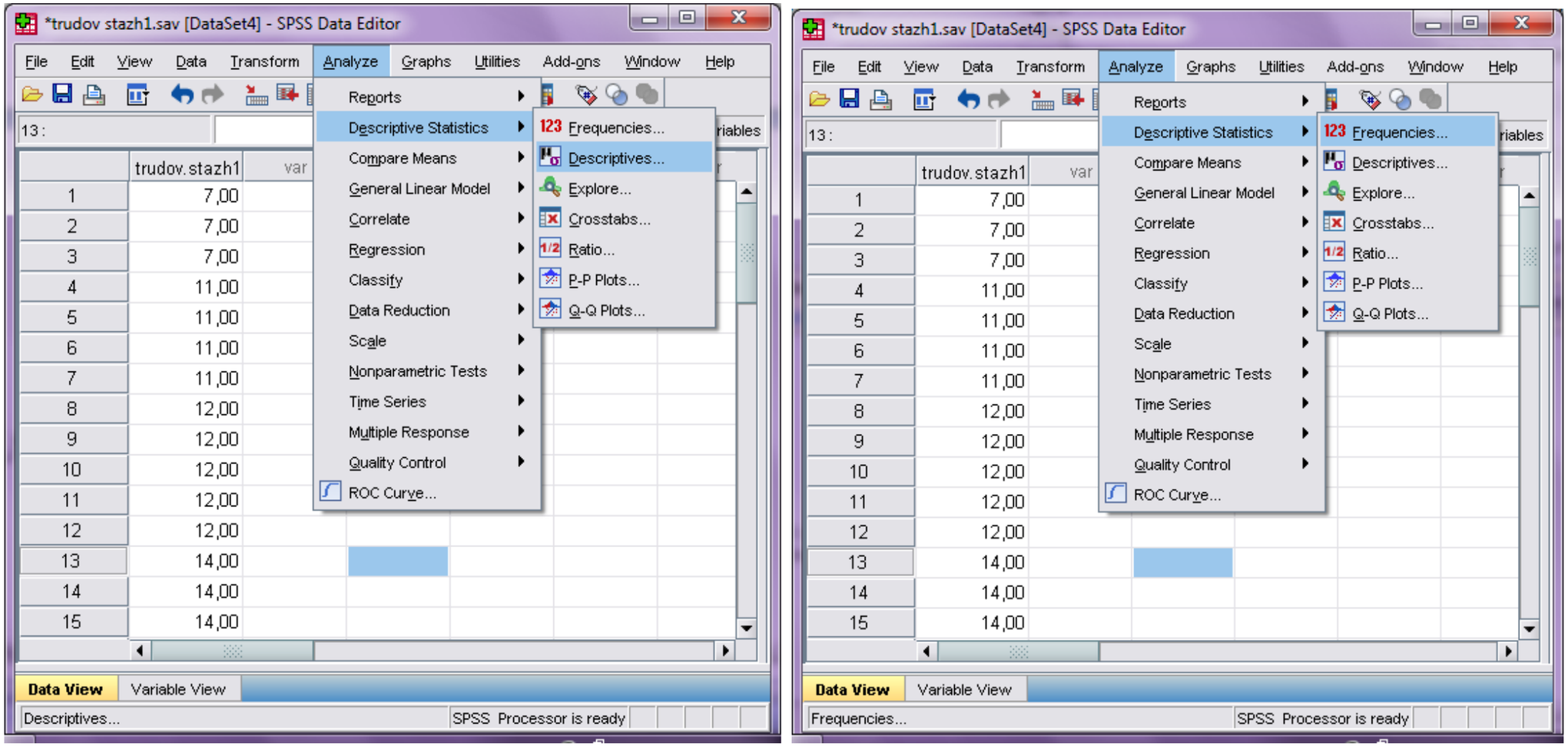
Mo = 14

$$s^2 = \frac{1}{n_1 + n_2 + \ldots + n_k - 1} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 =$$

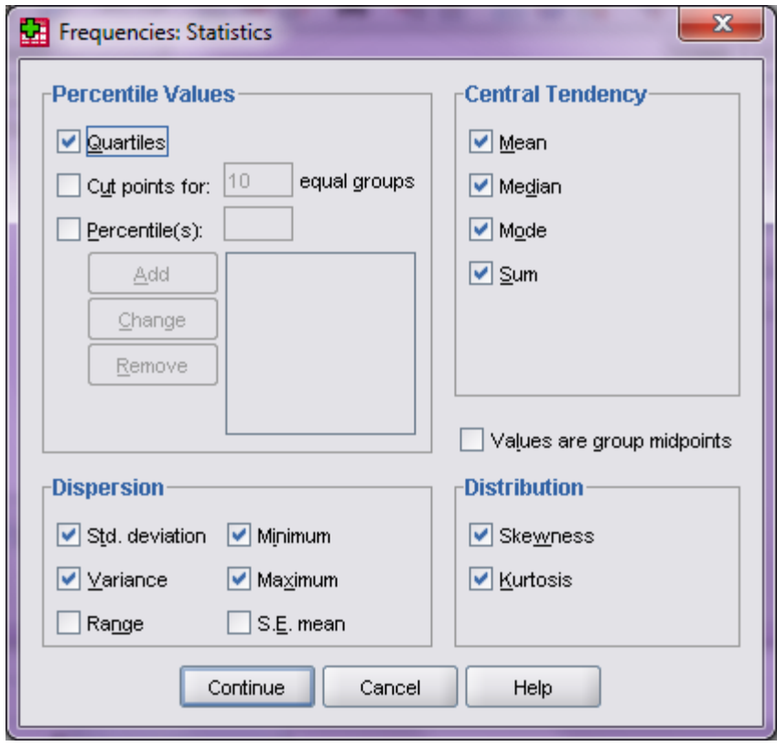
$$\frac{(7-12{,}41)^2.3 + (11-12{,}41)^2.4 + (12-12{,}41)^2.5 + (14-12{,}41)^2.6 + (15-12{,}41)^2.2 + (17-12{,}41)^2.2}{21}$$

$$s^2 = 7{,}968$$

$$s = \sqrt{s^2} = \sqrt{7{,}968} = 2{,}823$$

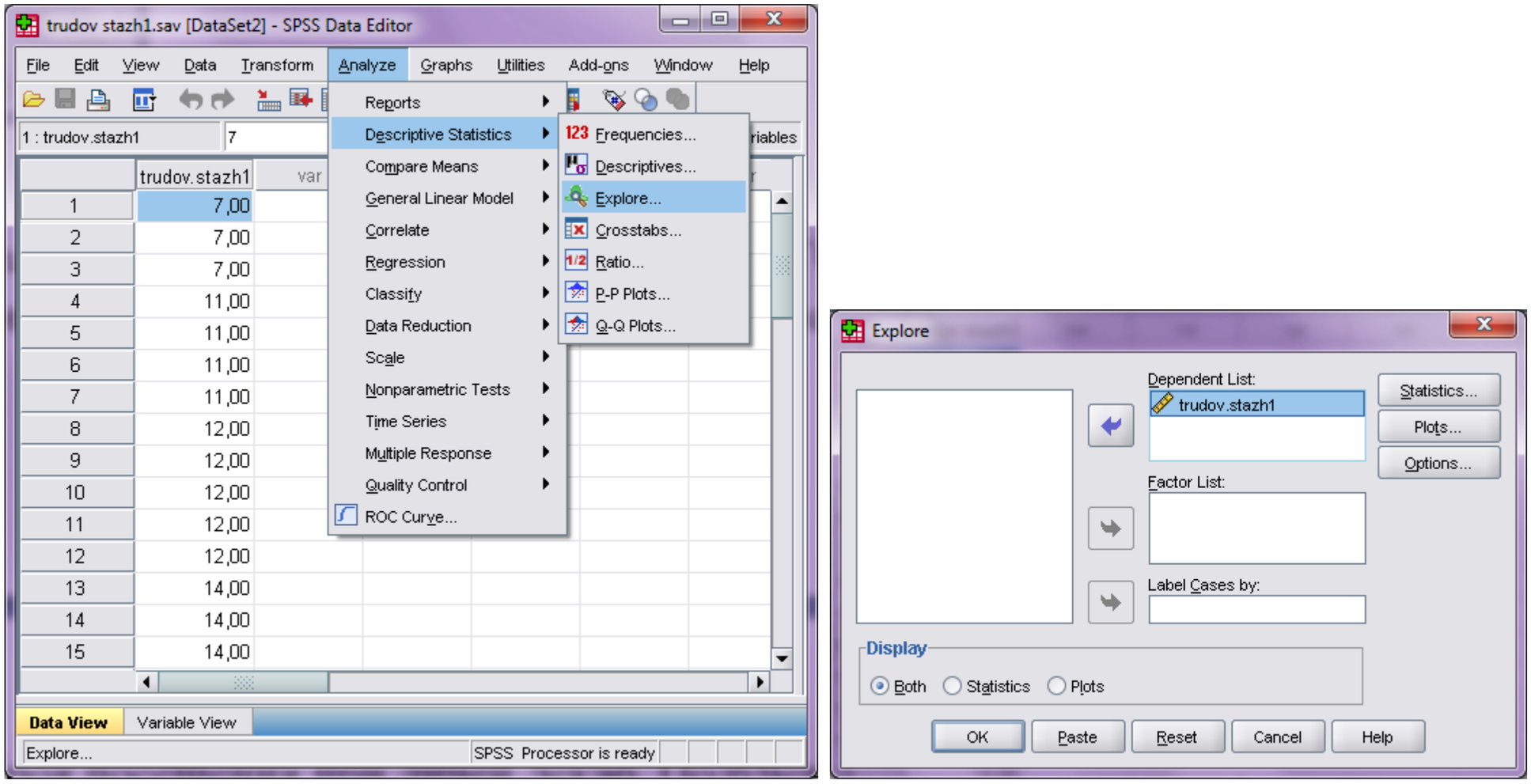

*trudov stazh1.sav [DataSet4] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Compare Means
General Linear Model
Correlate
Regression
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Multiple Response
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
trudov.stazh1
var
1 7,00
2 7,00
3 7,00
4 11,00
5 11,00
6 11,00
7 11,00
8 12,00
9 12,00
10 12,00
11 12,00
12 12,00
13 14,00
14 14,00
15 14,00
Data View
Variable View
Descriptives...
Frequencies...
SPSS Processor is ready

Фиг. 9.

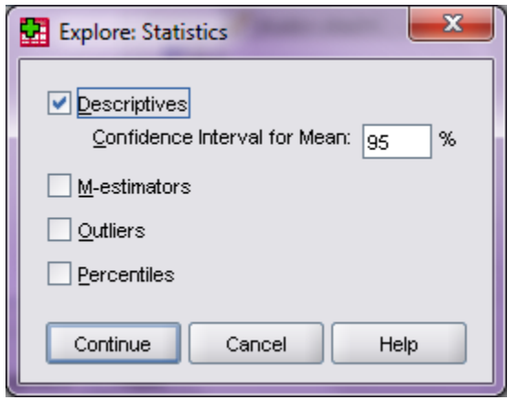


Фиг. 10.

Доверителен интервал за средно. Ако означим с μ средното на ГС, интервала (L_1,L_2) се нарича γ % доверетелен интервал за средното на генералната съвкупност, ако $P(L_1< \mu <L_2) = \gamma$.


trudov stazh1.sav [DataSet2] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : trudov.stazh1
7
trudov.stazh1
var
1 7,00
2 7,00
3 7,00
4 11,00
5 11,00
6 11,00
7 11,00
8 12,00
9 12,00
10 12,00
11 12,00
12 12,00
13 14,00
14 14,00
15 14,00
Reports
Descriptive Statistics
Compare Means
General Linear Model
Correlate
Regression
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Multiple Response
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
Data View
Variable View
Explore...
SPSS Processor is ready
Explore
Dependent List:
trudov.stazh1
Factor List:
Label Cases by:
Statistics...
Plots...
Options...
Display
Both
Statistics
Plots
OK
Paste
Reset
Cancel
Help

Фиг. 11.



Фиг. 12.

Задача 5. Изследва се броя посещения при личен лекар. Получена е следната извадка за 50 работни дни.

а) Да се напише вариационният ред и пресметнат размахът на извадката;

б) Да се състави статистическото разпределение на относителните честоти;

в) Да се начертае хистограма на честотите;

г) Да се пресметнат долният и горният квантил, медианата и модата.

д) Да се начертае бокс-плотът на разпределението.

е) Да се намерят извадъчните средна, дисперсия и средно квадратично отклонение;
ж) Да се намери 90% и 95% доверителен интервал за средния брой посещения при личният лекар. Как нивото на доверие влияе на доверителния интервал – с увеличаване нивото на доверие, нараства или намалява доверителният интервал?

1)

12	4	12	4	12	10	12	6	2	10
10	12	14	4	10	2	10	8	14	14
8	8	14	4	10	16	14	16	12	10
14	16	6	6	6	12	12	16	8	10
12	12	16	14	16	12	16	10	12	12

2)

16	14	15	20	16	11	13	2	14	6
8	18	0	16	8	18	16	10	18	17
14	18	10	19	12	2	10	6	10	14
13	18	16	7	8	7	7	13	15	10

14 8 20 2 18 8 12 14 18 12

Решете условията а) – ж) за задачи 6, 7 и 8.

Задача 6. В един магазин мандарините се продават в мрежи. Претеглени са 21 мрежи мандарини, избрани случайно. Известно е че данните са нормално разпределени и са дадени в таблицата.

3,3	3,5	3,9	4,0	3,2	3,5	3,8	4,1	3,9	4,2	4,1	3,8	3,6	3,2	3,3	3,6	3,7	3,5	3,4	3,5	3,9

Задача 7. За да планува бюджета за следващия месец собственик на малък хотел събира данни за платената цена от един човек в неговия хотел за дадения сезон. Направена е извадка от сметките на 20 клиента на ресторанта:

77 82 76 78 82 67 69 70 85 88 72 78 90 86 89 60 75 77 64 79

Задача 8. За да планира бюджета си собственик на малка фирма прави извадка от стойностите на печалбата от продажбата на 35 изделия:

0,60	1,00	0.78	0,70	0,90	0.88	0,50	1,08	0.20	0,90	0,60	0,80
0,45	0,45	0,68	0,44	0,78	0,65	0,75	0.60	0,50	0,66	0,58	0,64
0,90	0,55	0,74	0,75	0,65	0,30	0,95	0,70	0,45	0,90	0,80.	