Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ "МЭИ"

Институт информационных и вычислительных технологий

Кафедра математического и компьютерного моделирования

Отчёт по лабораторной работе №2 "Выборки и их представление"

Студент: Симаков А.М.
Преподаватель: Шевченко О.В.

Москва 2023

1 Основные понятия

Выборкой $x_1, ..., x_n$ объема n из совокупности, распределенной по $F(x)$, называется n независимых наблюдений над случайной величиной ξ с функцией распределения $F(x)$.

Вариационным рядом $x_1 \leq x_2 \leq ... \leq x_n$ называется выборка, записанная в порядке возрастания ее элементов.

Каждому наблюдению из выборки присвоим вероятность, равную $1/n$. Получим распределение, которое называют **эмпирическим**. Ему соответствует **функция эмпирического распределения**.

$$F_n^*(x) \equiv F_n^*(x; x_1, x_2, ..., x_n) = \frac{\mu_n(x)}{n},$$

где $\mu_n()$ - число членов выборки, меньших x.
Значение этой функции для статистики определяется следующим свойством (теорема Гливенко):

$$F_n^*(x) \to F(x) \text{ при } n \to \infty$$

Выборки больших объемов труднообозримы, поэтому разобьем диапазон значений выборки на равные интервалы и подсчитаем для каждого интервала **частоту** - количество наблюдений, попавших в него. Частоты, отнесенные к общему числу наблюдений n, называют **относительными частотами**. Графическое представление распределения частот по интервалам **гистограммой**. **Накопленной частотой** для данного интервала называют сумму частот данного интервала и всех тех, что левее него.
Числовые характеристики эмпирического распределения называются **выборочными характеристиками**: выборочные **среднее** (математическое ожидание), **дисперсия**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

выборочный **момент** порядка k:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$$

выборочные **квантили** ζ_p порядка p - корни уравнения

$$F(\zeta_p) = p,$$

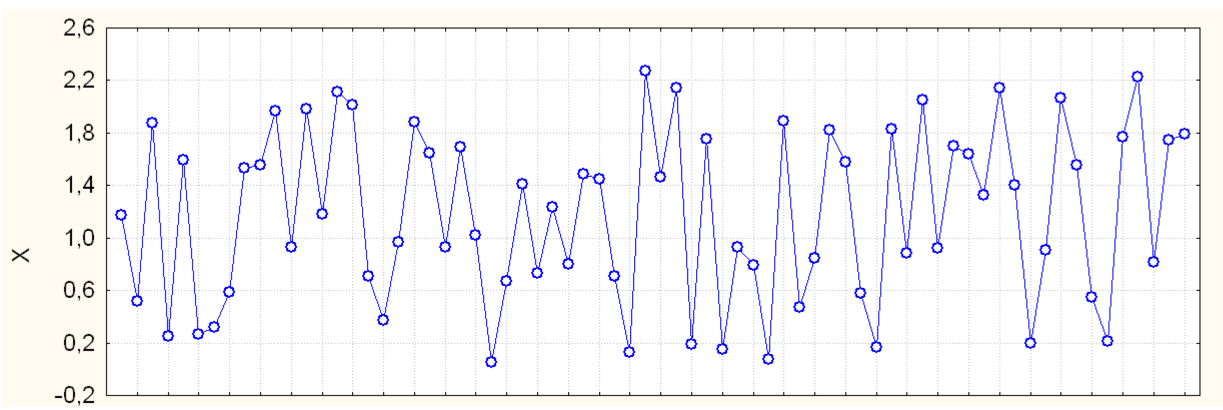
которыми являются члены вариационного ряда

$$\zeta_{(p)} = \xi_{[np]+1},$$

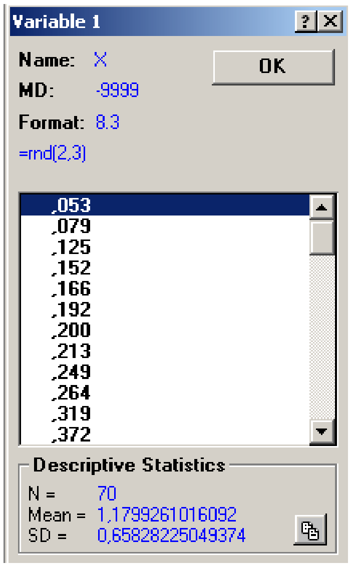
где $[np]$ означает целую часть np. Частным случаем ($p = 0.5$) является выборочная **медиана** - центральный член вариационного ряда. Значение выборочных характеристик состоит в том, что при $n \to \infty$ они стремятся к истинным значениям распределения $F(x)$.

Приведем с помощью пакета Statistica примеры.

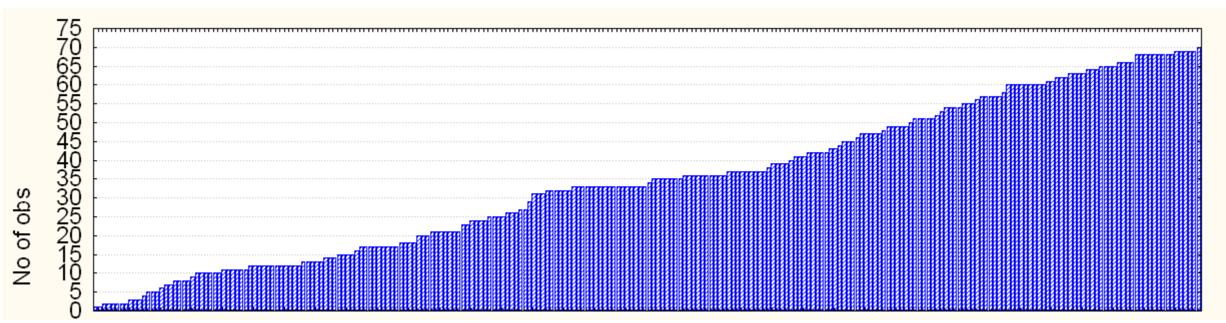
Сгенерируем выборку объёма $n = 70$ с распределением $\mathcal{R}[2,3]$ и представим её в графическом плане.



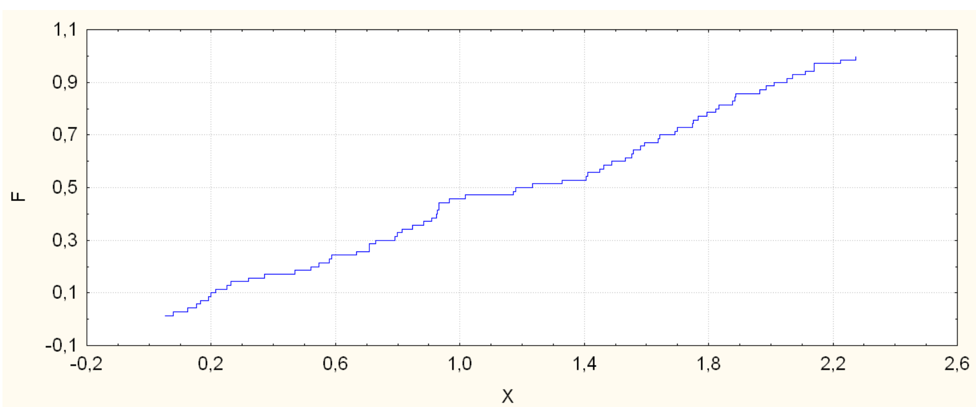
Построим **вариационный ряд** для данной выборки.



Построим функцию **эмпирического распределения** двумя способами.
1) Построение с помощью гистограмм



2) Построение с использованием новой переменной

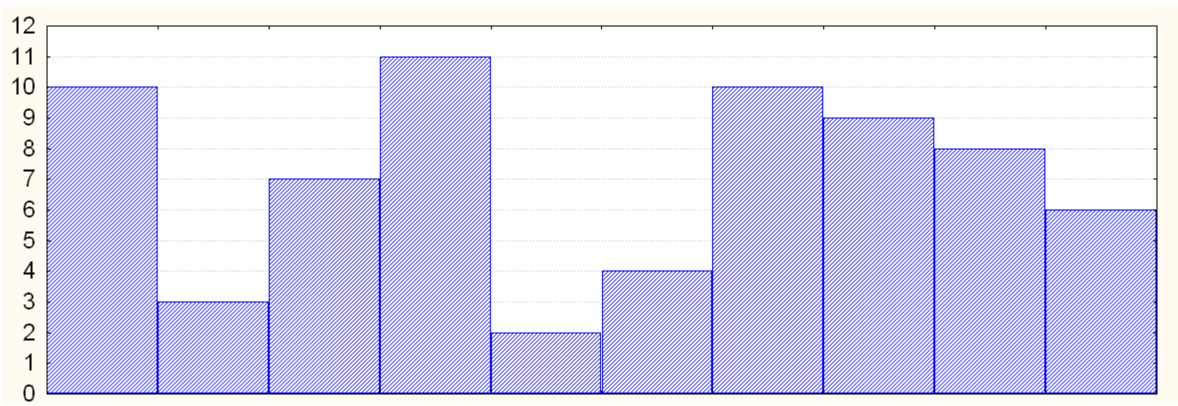


Группирование данных

Distribution: X: =rnd(2,3) (group_data.sta)

Continue...	Count	Cumul. Count	% of Non Missing	% of Selected	Cum.% of Non-Miss	100-%Non Missing	Cum.% of Selected	100-%of Selected
0,0000 < x <= ,5000	13	13	18,57143	18,57143	18,5714	100,0000	18,5714	100,0000
,5000 < x <= 1,0000	19	32	27,14286	27,14286	45,7143	81,4286	45,7143	81,4286
1,0000 < x <= 1,5000	10	42	14,28571	14,28571	60,0000	54,2857	60,0000	54,2857
1,5000 < x <= 2,0000	20	62	28,57143	28,57143	88,5714	40,0000	88,5714	40,0000
2,0000 < x <= 2,5000	8	70	11,42857	11,42857	100,0000	11,4286	100,0000	11,4286
Missing	0	70		0,00000			100,0000	0,0000
Not Selected	0	70						

Построение гистограмм частот



Выборочные характеристики

Descriptive Statistics (sample_char.sta)

Continue...	Valid N	Mean	Confid. -95,000%	Confid. 95,000	Sum	Minimum	Maximum	Range	Variance	Std.Dev.	Standard Error	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
X	70	1,179926	1,022964	1,336888	82,59483	,052925	2,272906	2,219980	,433336	,658282	,078680	-,130527	,286750	-1,25603	,566265

Описание двумерных выборок

Сгенерируем две выборки и построим диаграмму рассеяния. **не строитсся**

Выборочные характеристики

Descriptive Statistics (descr_of_two_dim.sta)

	Valid N	Mean	Confid. -95,000%	Confid. 95,000	Sum	Minimum	Maximum	Range	Variance	Std.Dev.	Standard Error	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
X	70	1,179926	1,022964	1,336888	82,59483	,052925	2,272906	2,219980	,433336	,658282	,078680	-,130527	,286750	-1,25603	,566265
Y	70	1,191013	1,039202	1,342823	83,37088	,069842	2,284417	2,214576	,405359	,636678	,076098	,022694	,286750	-1,10440	,566265

Определим корреляционную матрицу:

Correlations (descr_of_two_dim.sta)

Marked correlations are significant at p < ,05000
N=70 (Casewise deletion of missing data)

Variable	X	Y
X	1,00	-,09
Y	-,09	1,00

Двумерная гистограмма