

Изходен параметър Y	
1.098612289	
1.098612289	
1.093842598	

ТЕМА 13

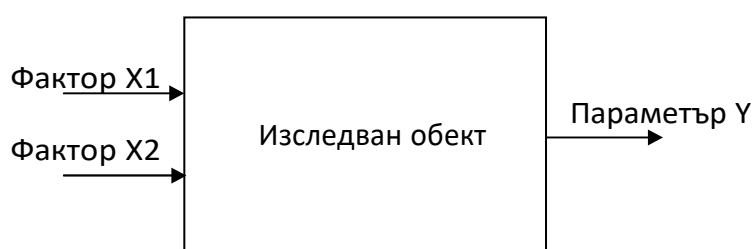
СТАТИСТИЧЕСКИ АНАЛИЗ НА ЕКСПЕРИМЕНТАЛНИ ДАННИ

Цел на упражнението

Да се направи статистически анализ на експериментални данни със средствата на MS EXCEL.

Теоретична част

Всеки изследван обект, представен като „черна кутия“, се характеризира със входни въздействия X и изходни реакции Y (фиг.1).



Фиг. 1. Изследван обект, представен като „черна кутия“

Входните въздействия X се представят с независими променливи. Наричат се **фактори** и определят поведението на обектите.

Изходните реакции Y се представят с изходни величини. Наричат се **параметри** и изразяват изменението на обектите под действието на факторите.

Всеки изследван обект може да бъде определен чрез факторите и параметрите му.

Връзките между изследваните параметри на обектите могат да се изследват със статистически модели. Тези модели се получават като резултат от една от основните стъпки на статистическото моделиране т.е. при избор на математически метод за статистическа обработка на извадката и получаване на статистически оценки за изследваните параметри.

Най-често използваните методи за статистическа обработка на данните са: корелационен анализ; регресионен анализ; дисперсионен анализ; Статистически оценки и проверка на х

Корелационният анализ изследва взаимовръзката между две променливи X и Y. При него се определят оценките на корелационните коефициенти и корелационната зависимост. Тя може да бъде:

- по форма - линейна или нелинейна;
- по характер - права (положителна) или обратна (отрицателна);
- според броя на едновременно действащите фактори - единична или множествена.

Основни величини в корелационния анализ са коефициентът на корелация и корелационното отношение. **Коефициентът на корелация** е число от интервала $[-1, +1]$, с което количествено се изразява съществуващата зависимост между две величини X и Y. От стойностите на корелационния коефициент се определя и степента на корелираност. Колкото абсолютната стойност на коефициента на корелация е по-близка до единица, толкова по-силна е зависимостта между компонентите X и Y.

Корелационното отношение R е количествена характеристика на корелационна зависимост. Приема стойности в интервала $[0,1]$. Ако стойността на корелационното

отношение е 0, то няма корелационна зависимост между величините Y и X . При $R = 1$ зависимостта е функционална, а при $0 < R < 1$ зависимостта е стохастична.

Дисперсионният анализ се прилага за установяване на относителното влияние на различни фактори върху стойностите на изходните характеристики. Обикновено се използва за сравнение на средните стойности на отделни слоеве на една извадка (еднофакторен анализ) или при изследване на диференциалните ефекти при два фактора (двуфакторен анализ).

При еднофакторния анализ се използва дисперсионен модел като математическо съотношение, при което всяка променлива се представя като сума от средна стойност и грешка.

При **двуфакторния дисперсионен анализ** се изследват два типа отношения:

- пресичане, когато двата фактора са представени чрез всички възможни комбинации за отделните нива;
- групиране, когато всяко отделно ниво на единия фактор участва в комбинации само с едно ниво на другия фактор.

Дисперсионният анализ е статистически метод за проверка на хипотези. Чрез тази проверка може да се прецени доколко влиянието на даден фактор или на група фактори е статистически значимо или не. Изследването на влиянието на факторите с помощта на оценките на дисперсиите е неговата същност. Ако оценките не се различават съществено се счита, че влиянието на фактора не е съществено. В противен случай се приема обратното твърдение (нулевата хипотеза H_0 се отхвърля).

Проверката на хипотези е процедура за приемане или отхвърляне на твърдение относно стойностите на един или няколко параметъра.

Нулевата хипотеза (H_0) гласи, че изследваният фактор (X_i) няма съществено значение за промените в изходния параметър (Y). **Алтернативната хипотеза (H_1)** гласи, че изследваният фактор има значимо влияние върху промените в изходния параметър.

Практическа част

За обекта на изследване от фиг. 1 са направени 1200 измервания на нивата на входните фактори X_1 и X_2 и изходния параметър Y . На всеки 10 измервания конфигурацията (вътрешната структура) на обекта се променя. Получените експериментални данни за записани във файл с име [regression.xls](#).

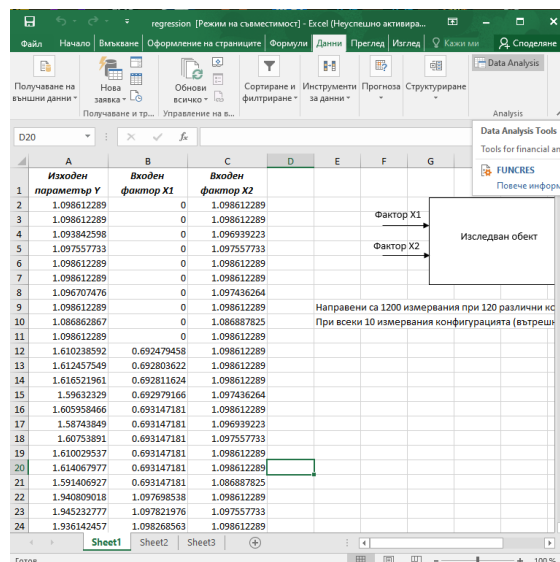
Да се провери какво е влиянието на факторите X_1 и X_2 върху изменението на изходния параметър Y . Да се даде отговор на въпроса: Влиянието на даден фактор X_i дали е съществено (значимо) за изменението на изходния параметър (Y)?

Да се провери значимостта на получения регресионен модел.

За решаване на поставената задача е необходимо да се извършат дисперсионен и корелационен анализи.

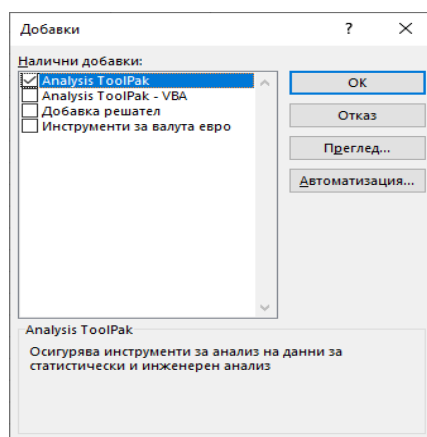
За да се извърши двуфакторен дисперсионен анализ със средствата на MS EXCEL е необходимо да се изпълнят следните действия:

1. От таба Data (Данни) в групата Analysis (Анализи) да се избере Data Analysis (фиг.2).



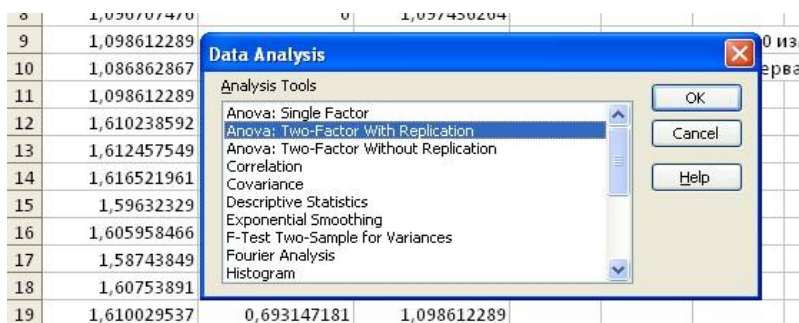
Фиг. 2. Избор на Data Analysis

Ако групата Analysis липсва, изберете File->Options->Add-Ins (Файл->Опции->Добавки) В полето Manage (Управление) изберете Excel Add-Ins (Добавки на Excel), а след това щракнете върху Go (Почни). В полето Добавки отметнете в квадратчето пред Analysis ToolPak и след потвърдете с ОК както е показано на фиг.3.



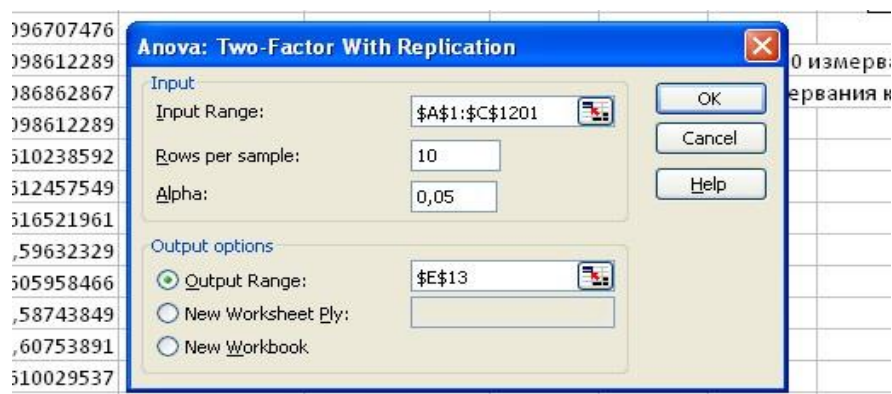
Фиг. 3. Добавяне на Analysis ToolPak

2. Изберете Anova: Two Factor With Replication и потвърдете с ОК



Фиг. 4. Избор на Anova: Two Factor With Replication

3. В полето Input Range: задайте областта от клетки A1:C1201 (фиг.5).
4. В полето Output Range: задайте клетката E13(фиг.5).
5. В полето Rows per sample: въведете 10. Потвърдете с ОК (фиг.5).



Фиг. 5. Настройки на полета

Крайният резултат от извършения анализ е представен на фиг.6:

	Total					
Count	1200	1200				
Sum	2536,23	1877,748				
Average	2,113525	1,56479				
Variance	0,625054	0,082341				
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample	422,6144	119	3,551381	3773,665	0	1,230346
Columns	180,666	1	180,666	191974	0	3,845767
Interaction	423,5195	119	3,558987	3781,747	0	1,230346
Within	2,032768	2160	0,000941			
Total	1028,833	2399				

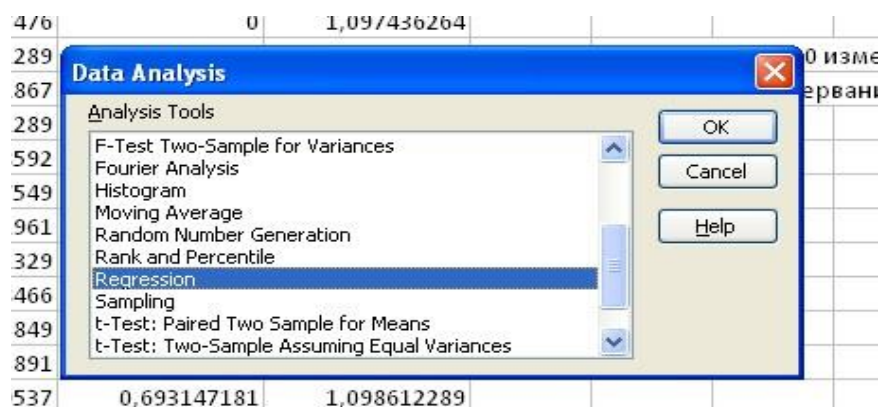
Фиг. 6. Получени резултати от анализа

На базата на получените резултати могат да се направят следните изводи:

- 1) $F > F_{crit}$, следователно нулевата хипотеза се отхвърля;
- 2) Влиянието на двата входни фактора X_1 и X_2 върху изходния параметър Y е значително.

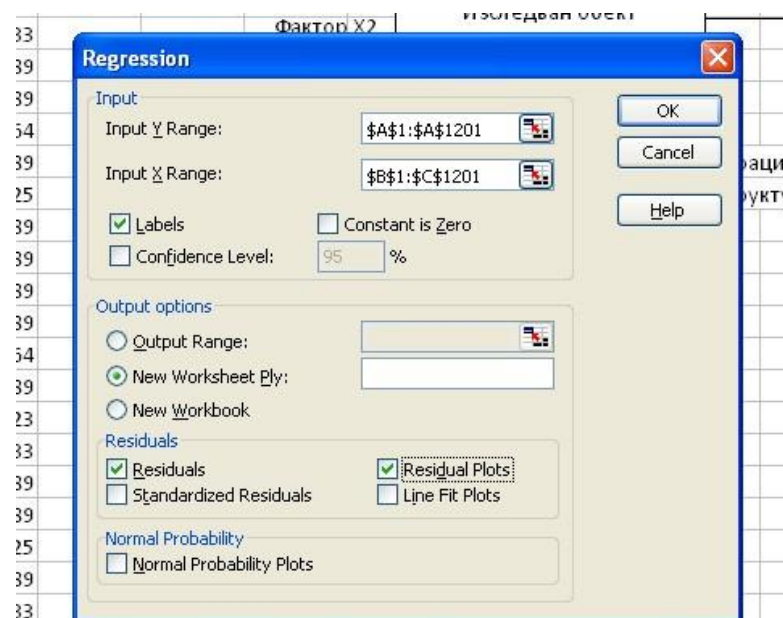
За да се намери каква точно е взаимовръзката между изходния параметър и входните фактори е необходимо да се направи регресионен анализ. За целта трябва да се изпълнят следните действия:

1. От менюто Tools да се избере Data Analysis (фиг.2)
2. Изберете Regression и потвърдете с ОК (фиг.7)



Фиг. 7. Избор на вид анализ

3. В полето Input Y Range: задайте областта от клетки A1:A1201 (фиг.8).
4. В полето Input X Range: задайте областта от клетки B1:C1201.
5. Изберете Labels.
6. Изберете Residuals и Residuals Plots. Потвърдете с ОК.



Фиг. 8. Настройки на полета за регресионен анализ

В нов работен лист се генерират резултатите, показани на фиг.9.

A	B	C	D	E	F	G	H	I
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0,984920759							
R Square	0,970068901							
Adjusted R Square	0,97001889							
Standard Error	0,110914804							
Observations	1200							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	477,2578692	238,6289	19397,42	0			
Residual	1197	14,72560628	0,012302					
Total	1199	491,9834755						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	0,551676757	0,019728365	27,96363	7,7E-133	0,512970734	0,5903828	0,5129707	0,5903828
Входен фактор X1	0,766278338	0,004051555	189,1319	0	0,758329398	0,7742273	0,7583294	0,7742273
Входен фактор X2	0,617515892	0,011162791	55,31913	0	0,595615079	0,6394167	0,5956151	0,6394167
RESIDUAL OUTPUT								
Observation	Исходен параметър	Residuals						
1	1,230087305	-0,131475016						
2	1,230087305	0,121475016						

Фиг. 9. Получени резултати от регресионен анализ

От получените резултати в частта *Regression Statistics* (фиг.9) се вижда, че коефициента на множествена корелация R (*Multiple R*) е равен на 0,9849. Този резултат ни дава основание за следния извод:

- 1) Между изходния параметър Y и факторите X1 и X2 имаме много висока степен на свързаност (98,5 %)

Значимост на получения регресионен модел.

За да се провери дали получения регресионен модел е достоверен (статистически значим) се проверява стойността на Significance F (в случая 0), фиг. 10. Ако тази стойност е по-малка от нивото на значимост (в случая 0,05), то получения модел е статистически значим. Ако Significance F е по-голяма от 0,05, то е по-добре да се потърси друг набор от входни фактори за съставяне на модела. Премахва се фактора с висока стойност на P-value (по-висока от 0,05) и регресионния анализ се повтаря докато Significance F стане по-малко от 0,05.

Всички P-values трябва да бъдат по-малки от 0,05. В този случай това е така (7,7E-133; 0 и 0), фиг. 10.

ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	477,2578692	238,6289	19397,42	0	
Residual	1197	14,72560628	0,012302			
Total	1199	491,9834755				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0,551676757	0,019728365	27,96363	7,7E-133	0,512970734	0,5903828
Входен фактор X1	0,766278338	0,004051555	189,1319	0	0,758329398	0,7742273
Входен фактор X2	0,617515892	0,011162791	55,31913	0	0,595615079	0,6394167

Фиг. 10. Част от резултатите, необходими за оценка на регресионния модел

От фиг. 11 се виждат стойностите на коефициентите (*Coefficients*) на регресионната зависимост. Последната може да се представи със следното уравнение (1)

$$Y = 0,5517 + 0,7663 * X1 + 0,6175 * X2 \quad (1)$$

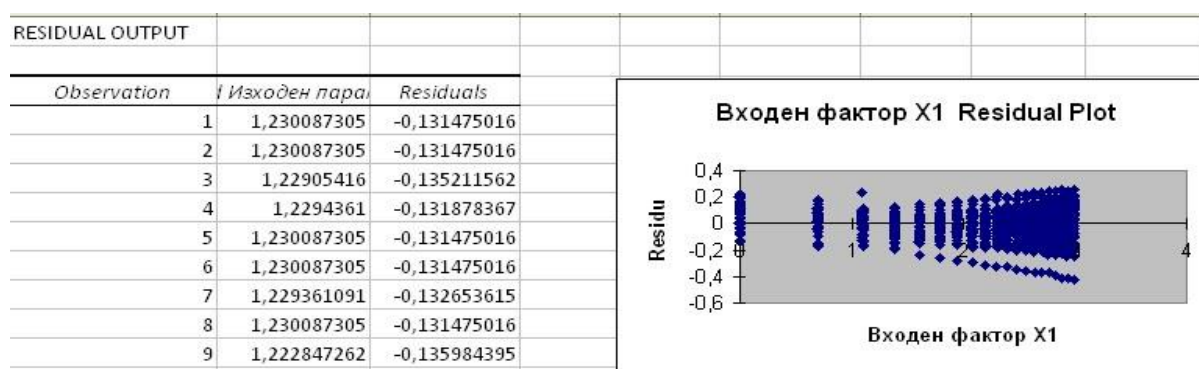
Всяко нарастване на X1 с единица води до нарастване на Y с 0,7636. Всяко нарастване на X2 с единица води до нарастване на Y с 0,6175 (фиг.11).

ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	477,2578692	238,6289	19397,42	0	
Residual	1197	14,72560628	0,012302			
Total	1199	491,9834755				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0,551676757	0,019728365	27,96363	7,7E-133	0,512970734	0,5903828
Входен фактор X1	0,766278338	0,004051555	189,1319	0	0,758329398	0,7742273
Входен фактор X2	0,617515892	0,011162791	55,31913	0	0,595615079	0,6394167

Фиг. 11. Коефициенти на регресионната зависимост

Така получения регресионен модел може да се използва за предварителна оценка на изходната реакция на изследвания обект при определени входни въздействия.

Остатъчните оценки (*Residuals*, фиг. 12) показват, каква е разликата между актуалните данни и изчислените посредством регресионния модел (регресионното уравнение).



Фиг. 12. Остатъчните оценки и диаграма на разсейването

Диаграмите показват разсейването на остатъчните оценки в зависимост от входните фактори X1 и X2.

Задачи за изпълнение

1. Запознайте се с теоретичната и практическа част на упражнението.
2. Решете поставената задача в практическата част. Реализираните дисперсионен и корелационен анализ като изпълните посочените действия в тази част.
3. Направете дисперсионен анализ на данните, записани в таблица Sheet2 на файла regression.xls.

A	B	C	D	E	F	G	H	I	J	K	L
Инженери	Медици	Икономисти									
42	69	35		Да се установи дали има значими различия в средните стойности на заплатите на хора с образование в различни области.							
53	54	40		Използвайте еднофакторен дисперсионен анализ "Anova: Single Factor"							
49	58	53		H0: $\mu_1 = \mu_2 = \mu_3$							
53	64	42		H1: поне една от областите има значима статистическа разлика, т.е. трите области не са статистически еднакви							
43	64	50		След като получите резултатите от анализа проверете стойностите, които стоят в колоната Average, там е записана средната заплата според полученото образование. За да се определи дали между тези средни стойности има значителна статистическа разлика трябва да се погледне стойността в колоната P-Value. Ако тя е по-малка от нивото на значимост (0,05 в случая), може да се каже че има значителни разлики в заплатите в зависимост от получената образователна степен.							
44	55	39		Освен това ако $F > F_{crit}$, нулевата хипотеза се отхвърля, което ще означава че значенията на трите области не са статистически еднакви и поне едно от значенията е различно.							
45	56	55									
52		39									
54		40									

4. Направете регресионен анализ на данните, записани в таблица Sheet3 на файла regression.xls.

A	B	C	D	E	F	G	H	I	J	K	L
Продадено количество	Ед. цена	Изразходвани средства за реклама									
8500	\$2	\$2 800		Има ли връзка и каква е тя между продадените количества (изходния параметър) и ед. цена и средствата за реклама (входните фактори), т.е. може ли да се определи продаденото количество, ако се знае цената и средствата за реклама?							
4700	\$5	\$200									
5800	\$3	\$400									
7400	\$2	\$500									
6200	\$5	\$3 200									
7300	\$3	\$1 800									
5600	\$4	\$900									

Въпроси и задачи за самоподготовка

1. С какво се характеризират изследваните обекти?
2. Кои са факторите и кои са параметрите на изследваните обекти?
3. Какви са задачите на статистическия анализ?
4. Какво е характерно за корелационния анализ?
5. Какво е характерно за дисперсионния анализ?
6. Какво наричаме корелационно отношение. Какви стойности може да приема?
7. Какво е коефициентът на корелация? За какво се използва? Какви стойности може да приема?