Статистика

Павлина Йорданова

ШУ “Епископ Константин Преславски”

Литература

Георги Мишев, Статистика за икономисти, ИК Люрен, София, 1993.

Екатерина Тошева, Статистическо изследване на зависимости - методическо ръководство, УНСС, София, 2012.

Павлина Йорданова, Евелина Велева, Статистическо моделиране на вероятностни разпределения с Excel, Университетско издателство „Епископ Константин Преславски", Шумен, 2017.

Серафим Петров, Снежана Велева-Стефанова, Обща теория на статистиката, УИ "Васил Априлов" , Габрово, 2001.

Конспект по Статистика

1. *Статистическо изучаване. Основни статистически понятия.*
2. *Статистическа групировка. Начини за представяне на информацията от наблюдението. Кръстосани таблици и статистически редове. Статистически графични изображения.*
3. *Статистически величини. Същност. Абсолютни и относителни статистически величини. Средни величини.*
4. *Статистическа вариация. Нормална крива. Моменти, асиметрия, ексцес.*
5. *Емпирична функция на разпределение и квантили.*

6. Графика с мустачки. Силно отличаващи се наблюдения.

7. Крива на Лоренц.

8. Елементи от Теорията на вероятностите. Опит. Събития. Вероятност.

9. Условна вероятност. Формула за умножение на вероятностите. Независимост и корелация между събития. Формула за пълната вероятност. Формула на Бейс.

10. Дискретни вероятностни разпределения. Числови характеристики. Примери.

11. Абсолютно непрекъснати вероятностни разпределения. Числови характеристики. Примери.

12. Гранични теореми, намиращи приложение в статистиката.

13. Графични методи за определяне на типа на разпределението на извадката. PP-plot и QQ-plot.

14. Статистическа оценка на параметри. Точкови оценки.

15. Интервални оценки. Доверителен интервал за средното на наблюдаваната случайна величина.

16. Доверителен интервал за относителен дял в генерална съвкупност.

17. Доверителен интервал за единична стойност на нормално разпределена случайна величина.

18. Статистическа проверка на хипотези. Общи сведения Проверка на хипотези за параметрите на една генерална съвкупност.

19. Проверка на хипотези за равенство между параметрите на две генерални съвкупности.

20. Проверка на непараметрични хипотези. $\chi 2$ – критерий на съгласие.

21. Еднофакторен дисперсионен анализ.

22. Корелационен анализ. Основни понятия.

23. Регресионен анализ.

24. Анализ на динамични редове. Описателни характеристики на динамичните редове.

25. Методи и модели за анализ на тенденцията в развитието.

26. Статистически анализ на сезонни колебания.

21. Еднофакторен дисперсионен анализ.

1. Приложение. Дисперсионният анализ се прилага, когато се интересуваме дали влиянието на един или няколко неметрирани фактор признаци върху друг метриран признак, наречен резултативен е статистически значимо. В зависимост от броя на фактор признаците имаме еднофакторен, двуфакторен и т.н., многофакторен дисперсионен анализ. Тук ще разгледаме само случая с един фактор-признак. По същество това е проверка на хипотези за равенство между средните на две или повече извадки от нормално разпределени и независими съвкупности при предположение, че дисперсиите им са равни. Използва се обикновено когато значенията на фактор-признака са повече от две, в противен случай бихме могли да използваме по-кратката проверка на хипотези за равенство между средни на две извадки. Задачата се свежда до проверка на хипотези за равенство между две дисперсии (междугрупова и вътрешногрупова). Методологията му е разработена от Р. Фишер.

Тъй като реализацията му е свързана с много пресмятания, обикновено тя се извършва с помощта на компютър. Например със Statistica, Excel, SPSS или др.

2. Постановка на задачата. *Нека наблюдаваме два признака* ξ *и* η *върху n статистически единици. Целта ни ще е да отговорим на въпроса дали влиянието на признака* ξ*, който не е задължително да е метриран и е с възможни значения* $X_1, X_2, \ldots, X_k$*, върху значенията на метрирания признак* η *е статистически значимо. Да предположим, че извадката е от нормално разпределена съвкупност и резултатите от наблюдението са дадени в Табл. 1. Наблюденията върху* η *в случаите, когато* $\xi = X_i$ *са* n_i *на брой и са означени с* $y_{in_i}, i = 1, 2, \ldots, k.$

ξ	η	k *Общо:*
X_1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	n_1
...	...	...
X_k	$y_{11}, y_{12}, \ldots, y_{kn_k}$	n_k
Общо:		n

Приемаме, че извадките в групите са независими. Да означим средната в i – тата група с

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}, \qquad i = 1,\ 2,\ ...,\ k.$$

Избираме ниво на съгласие α. Проверяваме хипотезата

H_0 : $E(\eta|\xi = X_1) = E(\eta|\xi = X_2) = ... = E(\eta|\xi = X_k)$, т.е. отклоненията между средните в различните групи се дължат на случайни, кратко временно действащи фактори, т.е. влиянието на фактор-признака върху резултативния признак не е статистически значимо.

Алтернативата е

H_1 : Поне две от средните $E(\eta|\xi = X_1)$, $E(\eta|\xi = X_2)$, …, $E(\eta|\xi = X_k)$, са различни, т.е. отклоненията между поне две от средните в различните групи се дължат на системни фактори, т.е. влиянието на фактор-признака върху резултативния признак е статистически значимо.

Избираме риск за грешка от първи род $\alpha \in (0, 1)$.

Като критерий за проверка на тези хипотези се използва отношението на междугруповата и вътрешногруповата дисперсии. За да ги дефинираме се нуждаем от следните понятия.

Обща девиация (отклонение, обща сума от квадратите SS_o) се нарича сумата от квадратите на отклоненията на всичките n измерени значения на метрирания признак от тяхната средна аритметична. Ще я означаваме с SS_0. Т.е. ако общата средна е

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{n}$$

тогава

$$SS_o = \sum_{i=1}^{k}\sum_{j=1}^{n_k} (y_{ij} - \bar{\bar{y}})^2.$$

Тя измерва разпръснатостта на единиците около общата средна. Има n – 1 степени на свобода.

Вътрешногрупова девиация (вътрешногрупова сума от квадратите) се нарича сумата от квадратите на отклоненията на всичките n измерени значения на метрирания признак от тяхната средна аритметична в съответната група. Ще я означаваме с SS_B. Т.е

$$SS_B = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Тази девиация има n – k степени на свобода.

Междугрупова девиация (междугрупова сума от квадратите) се нарича сумата от квадратите на отклоненията на средните аритметични в групите от общата средна аритметична претеглени с броевете на наблюденията в групите. Ще я означаваме с SS_M и

$$SS_M = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{\bar{y}})^2.$$

Тя има k - 1 степени на свобода.

Общата девиация е сума от вътрешногруповата и междугруповата девиации. Същото съотношение, както се вижда от по-горните разсъждения, е в сила и за степените им на свобода.

Общата девиация е сума от вътрешногруповата и междугруповата девиации. Същото съотношение, както се вижда от по-горните разсъждения, е в сила и за степените им на свобода. Т.е.

$$SS_0 = SS_B + SS_M,$$

$$n - 1 = n - k + k - 1.$$

Като разделим девиациите на степените им на свобода получаваме оценки за съответните дисперсии. Т.е.

вътрешногрупова дисперсия *ще наричаме* $\dot{S}_B^2 = \frac{SS_B}{n-k}$.

Междугрупова дисперсия *ще наричаме* $\dot{S}_M^2 = \frac{SS_M}{k-1}$.

Вече сме готови да построим критичната област за нулевата хипотеза. Тя е

$$W_\alpha = \left\{ (x_1,...,x_n) \in \Omega : \frac{\dot{S}_M^2}{\dot{S}_B^2} \geq x_{1-\alpha,F(k-1,n-k)} \right\}.$$

Случайната величина, която съответства на $\frac{\dot{S}_M^2}{\dot{S}_B^2}$ *има F разпределение с k – 1 степени на свобода на числителя и n – k степени на свобода на знаменателя. Поради това константата* $x_{1-\alpha,F(k-1,n-k)}$ *е 1-*α *квантила на това разпределение.*

Както и при проверката на хипотези за равенство между дисперсиите с критерия на Фишер, така и тук, критичната област може да се трансформира с еквивалентни преобразования, така че оценката на дисперсията от числителя да е по-голям от тази в знаменателя. Т.е. ако оценката на вътрешногруповата дисперсия е по-голяма от тази на междугруповата. Тогава

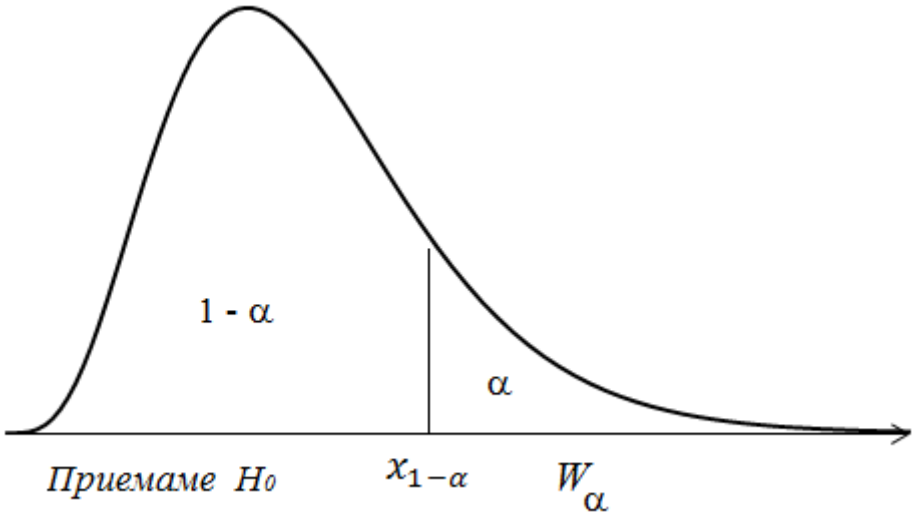
$$W_\alpha = \left\{ (x_1,...,x_n) \in \Omega : \frac{\dot{S}_B^2}{\dot{S}_M^2} < x_{1-\alpha,F(n-k,k-1)} \right\}, \quad x_{1-\alpha,F(n-k,k-1)} = \frac{1}{x_{1-\alpha,F(k-1,n-k)}}.$$

$x_{1-\alpha,F(n-k,k-1)}$ *е 1-*α *квантилът на F (n – k; k-1) разпределението.*

На следващата фигура е скицирана плътността на F разпределението с k – 1 степени на свобода на числителя и n – k степени на свобода на знаменателя, заедно с критичната област за нулевата хипотеза

$$W_\alpha = \{(X_1, X_2, \ldots, X_n) \in \Omega: \frac{\dot{S}_M^2}{\dot{S}_B^2} \geq x_{1-\alpha, F(k-1;n-k)}\},$$

където $x_{1-\alpha, F(k-1;n-k)}$ *е* $1 - \alpha$ *квантилът на същото разпределение.*



Формулата

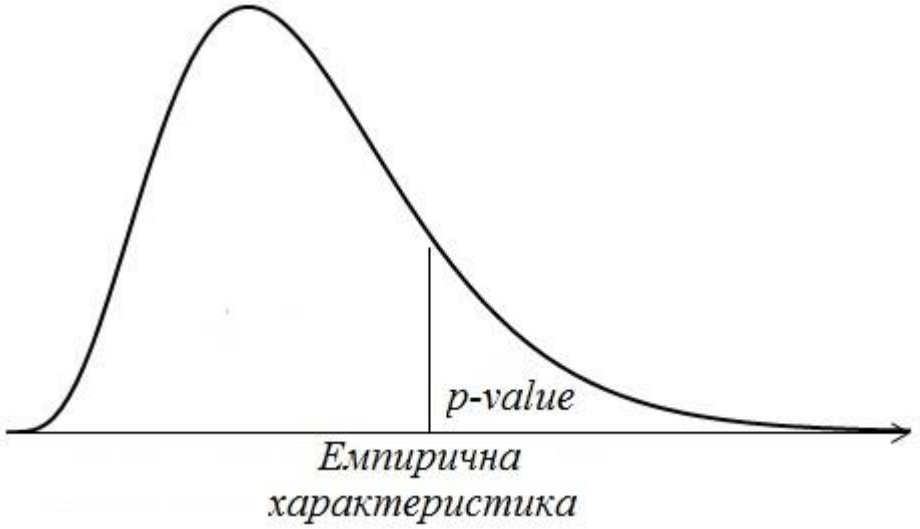
=F.INV(1-α; k – 1;n – k)

връща $x_{1-\alpha, F(k-1;n-k)}$.

Ако емпиричната характеристика е по-малка от този квантил, значи извадката не е в кр. обл. за H_0 *и нямаме основание да я отхвърлим.*

Т.е. влиянието на фактор-признака върху резултативния признак не е статистически значимо. Обратно: ако тя е по-голяма от $x_{1-\alpha, F(k-1;n-k)}$, *значи извадката е в кр. обл. за* H_0. *В този случай отхвърляме* H_0 *и приемаме* H_1. *Т.е. влиянието на фактор-признака върху резултативния признак е статистически значимо.*

Връзката между p-value, емпирична характеристика и графиката на същата плътност на F разпределението с k – 1 степени на свобода на числителя и n – k степени на свобода на знаменателя е изобразена на фигурата в дясно.



Функцията

=1 - F.DIST(x ; k -1; n – k; 1), с параметър x, равен на емпиричната характеристика $\frac{\dot{S}_M^2}{\dot{S}_B^2}$*, пресмята p-value = P(*ξ *> x), където*

$\xi \in F(k-1;\ n\ -k)$*. Сравняваме с предната графика и виждаме, че ако p-value* $> \alpha$*, значи извадката не е в критичната област за* H_0 *и нямаме основание да я отхвърлим. Обратно: ако тази стойност е по-малка от* α*, значи извадката е в критичната област за* H_0*. В този случай отхвърляме* H_0 *и приемаме* H_1*. Т.е. влиянието на фактор-признака върху резултативния признак е статистически значимо.*

В следващия пример ще обясним как може да ни бъде полезен Excel в случая.

	Вид на гумите			
Пробег в хиляди км.	A	B	C	D
	6,52	6,40	7,38	6,10
	8,72	8,00	8,38	6,72
	10,82	7,00	8,98	7,24
	9,92	8,40	8,78	9,68
	5,02	5,70	7,68	6,90
	11,42	10,00	9,28	11,73
	8,52	9,90	9,68	7,72
	8,02	5,20	5,68	4,97
	6,42	11,30	10,38	12,40
	9,22	9,50	11,68	8,41
	7,72	8,00	9,18	7,31
		9,40	7,48	7,71
		8,30		6,45
				8,50
				8,60
				8,70
				9,80

Пример 1. Завод произвежда 4 вида автомобилни гуми. Наблюдавани са 53 от тях. В таблицата в ляво е даден пробега им в хиляди км. до момента на пълното им износване, поотделно за четирите вида.

а) С различни цветове начертайте данните

на една и съща координатна система. Само като наблюдавате графиката оформете предварително ваше предположение за резултата от подточка г).

б) Начертайте 4 успоредни box-plot. Сравнете ги и направете извод.

в) Като използвате четири pp-plot проверете дали наблюдаваните четири условни разпределения са нормални (като пренебрегнете факта, че наблюденията са прекалено малко на брой за да получите правилно заключение).

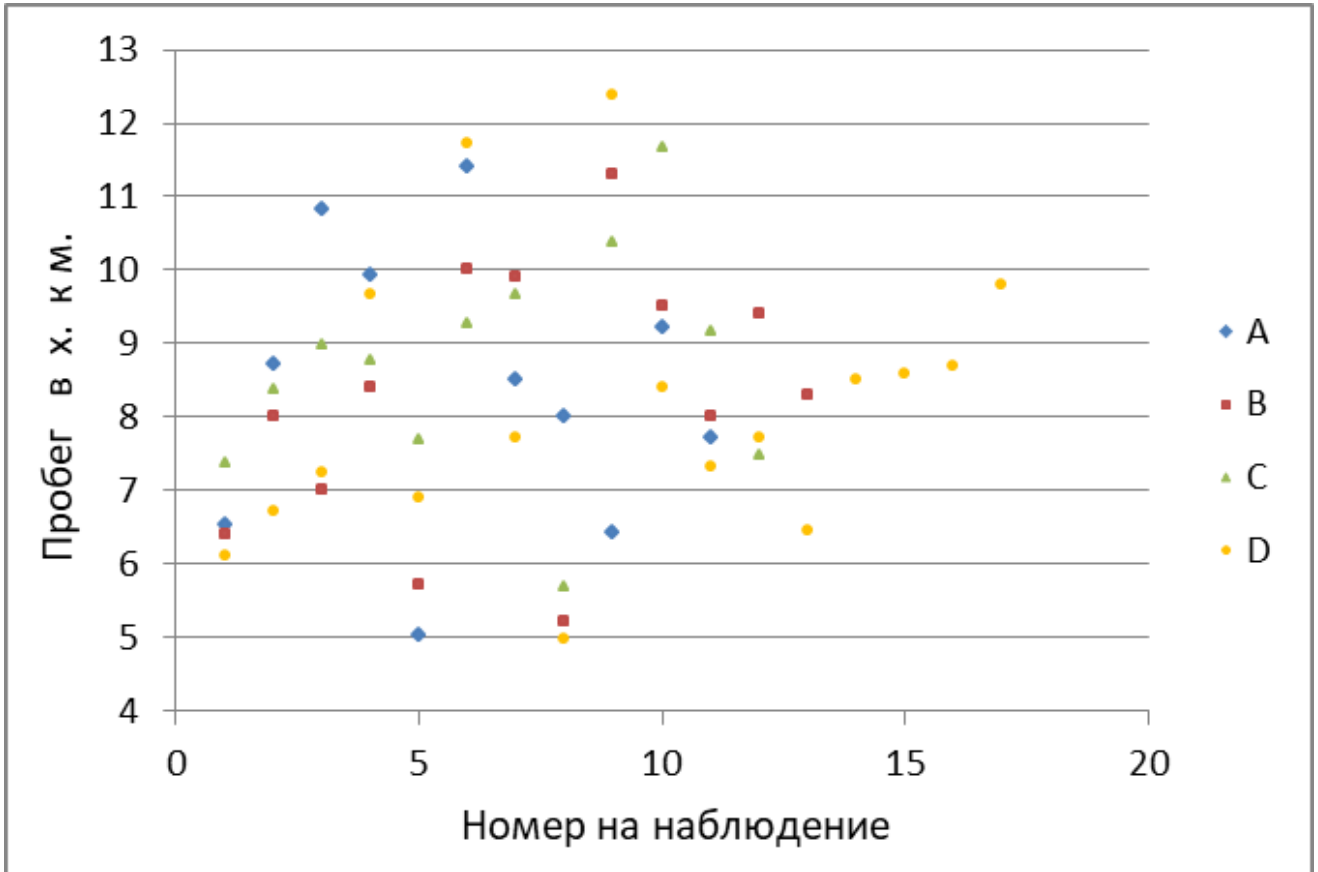
г) С риск за грешка от първи род $\alpha = 0.05$, проверете хипотезата, че вида на гумите не е статистически значим за пробега им.

Пример 1. Завод произвежда 4 вида автомобил-ни гуми. Наблюдавани са 53 от тях. В таблицата по-долу е даден пробега им в хиляди км. до момента на пълното им износване, поотделно за четирите вида.

№	A	B	C	D
1	6,52	6,40	7,38	6,10
2	8,72	8,00	8,38	6,72
3	10,82	7,00	8,98	7,24
4	9,92	8,40	8,78	9,68
5	5,02	5,70	7,68	6,90
6	11,42	10,00	9,28	11,73
7	8,52	9,90	9,68	7,72
8	8,02	5,20	5,68	4,97
9	6,42	11,30	10,38	12,40
10	9,22	9,50	11,68	8,41
11	7,72	8,00	9,18	7,31
12		9,40	7,48	7,71
13		8,30		6,45
14				8,50
15				8,60
16				8,70
17				9,80

а) С различни цветове начертайте данните на една и съща координатна система. Само като наблю-давате графиката оформете предварително ваше предположение за резултата от подточка г).

Решение: Insert -> Scatter



б) Начертайте 4 успоредни box-plot. Сравнете ги и направете извод.

Решение: Първо попълваме таблицата.

Характе-ристики	A			B			C			D		
	Стойности	Краища	Нараст-ване на кр.	Стойности	Краища	Нараст-ване на кр.	Стойности	Краища	Нараст-ване на кр.	Стойности	Краища	Нараст-ване на кр.
Min	5,02	3,45	3,45	5,20	3,25	3,25	5,68	5,01	5,01	4,97	4,20	4,20
Q1	7,12	7,12	3,68	7,00	7,00	3,75	7,63	7,63	2,63	6,90	6,90	2,70
Q2	8,52	8,52	1,40	8,30	8,30	1,30	8,88	8,88	1,25	7,72	7,72	0,82
Q3	9,57	9,57	1,05	9,50	9,50	1,20	9,38	9,38	0,50	8,70	8,70	0,98
Max	11,42	13,25	3,68	11,30	13,25	3,75	11,68	12,01	2,63	12,40	11,40	2,70

Като колонката с краища съдържа

$$Q1 - 1.5(Q3-Q1);\ Q1;\ Q2;\ Q3;\ Q3+1.5(Q3-Q1)$$

Таблицата попълваме например по следния начин. Функцията QUARTILE.INC с първи параметър колонката от данните, и втори параметър 0- за минимума, 1-за първия квартил, 2-за медианата, която съвпада с втория квартил, 3-за третия квартил и 4-за максимума. Т.е. ако буквата A от предната таблица ми е в клетка H2 функцията

=QUARTILE.INC($H$3:$H$13;0)

намира минималният пробег на гума от тип A.

б) Начертайте 4 успоредни box-plot. Сравнете ги и направете извод.

Решение: Първо попълваме таблицата.

Характе-ристики	A			B			C			D		
	Стойности	Краища	Нараст-ване на кр.	Стойности	Краища	Нараст-ване на кр.	Стойности	Краища	Нараст-ване на кр.	Стойности	Краища	Нараст-ване на кр.
Min	5,02	3,45	3,45	5,20	3,25	3,25	5,68	5,01	5,01	4,97	4,20	4,20
Q1	7,12	7,12	3,68	7,00	7,00	3,75	7,63	7,63	2,63	6,90	6,90	2,70
Q2	8,52	8,52	1,40	8,30	8,30	1,30	8,88	8,88	1,25	7,72	7,72	0,82
Q3	9,57	9,57	1,05	9,50	9,50	1,20	9,38	9,38	0,50	8,70	8,70	0,98
Max	11,42	13,25	3,68	11,30	13,25	3,75	11,68	12,01	2,63	12,40	11,40	2,70

Маркираме колонките с нарастванията на краищата на графиките с мустачки и даваме

Insert -> Bar

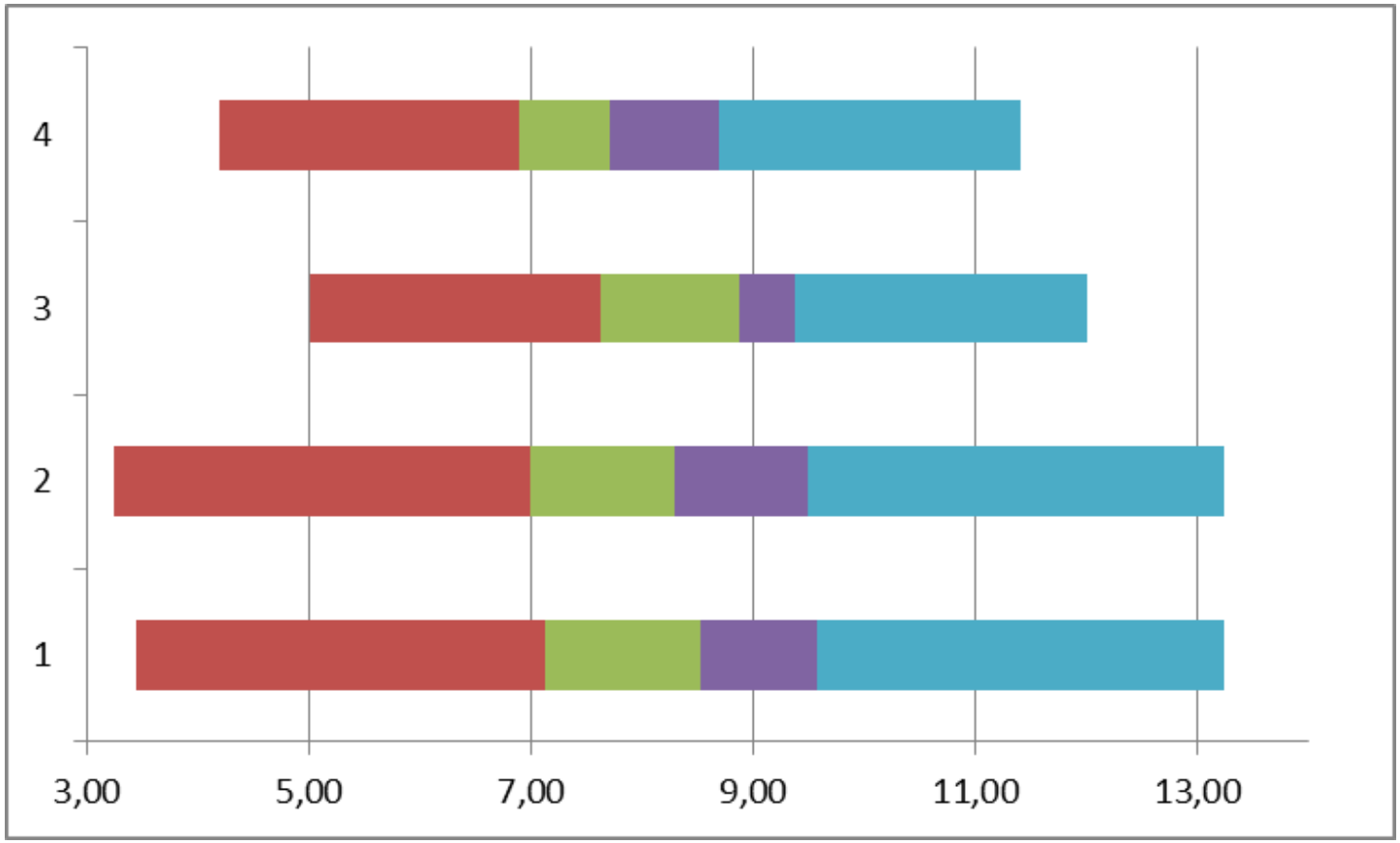
После от менюто Design -> Change row/Column

Изтриваме появилата се легенда.

Обезцветяваме най-левите правоъгълници на графиката.

Форматираме първата ос да започне от 3 и до свърши в 14.

Получаваме следващата фигурата.

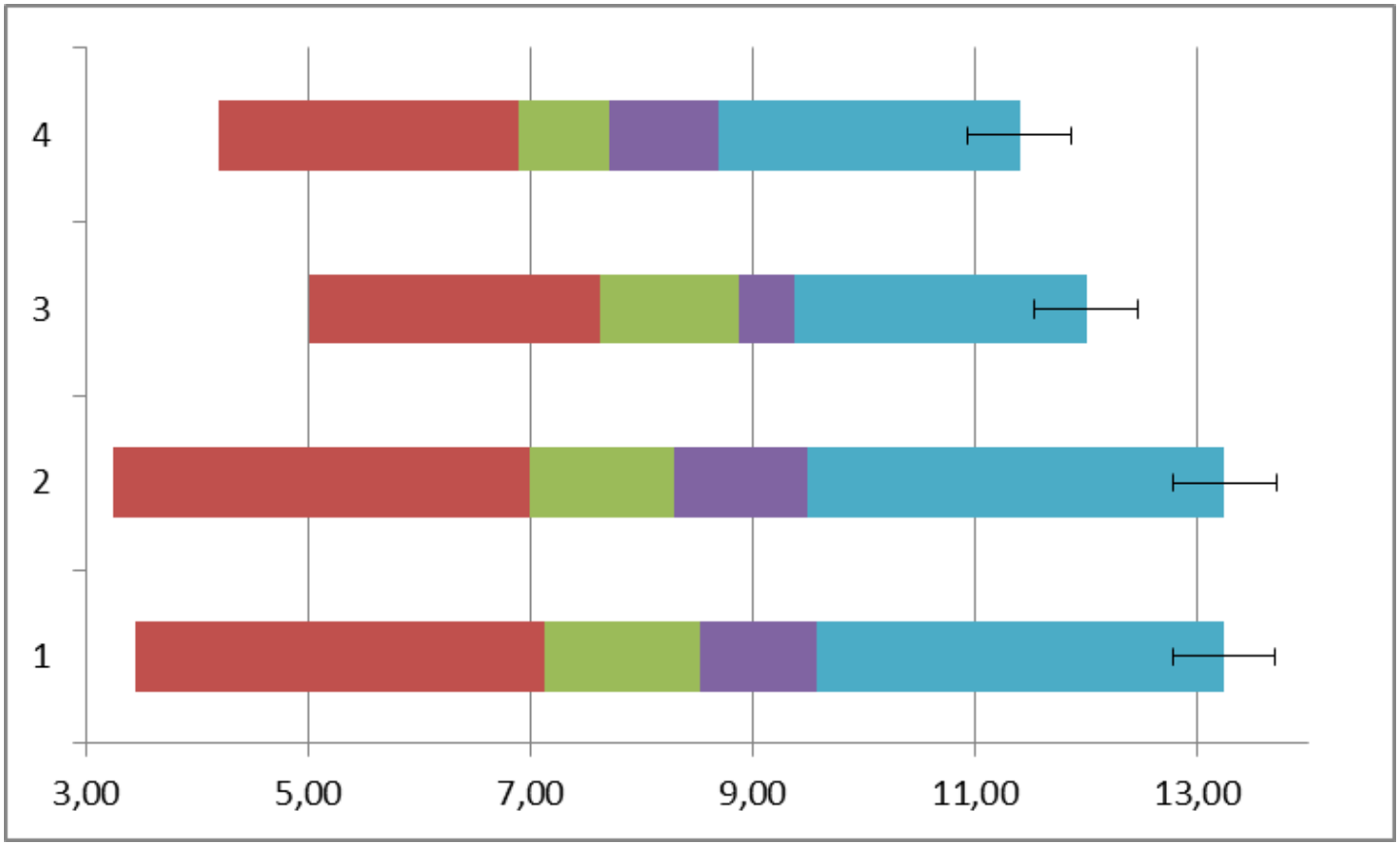


Още, обаче не сме готови.

Маркираме най-десните правоъгълници и от менюто

Layout -> Error bar -> Error bars with standard error

Получаваме следващата графика.



Все още не сме готови. Остана ни да форматираме мустачките.

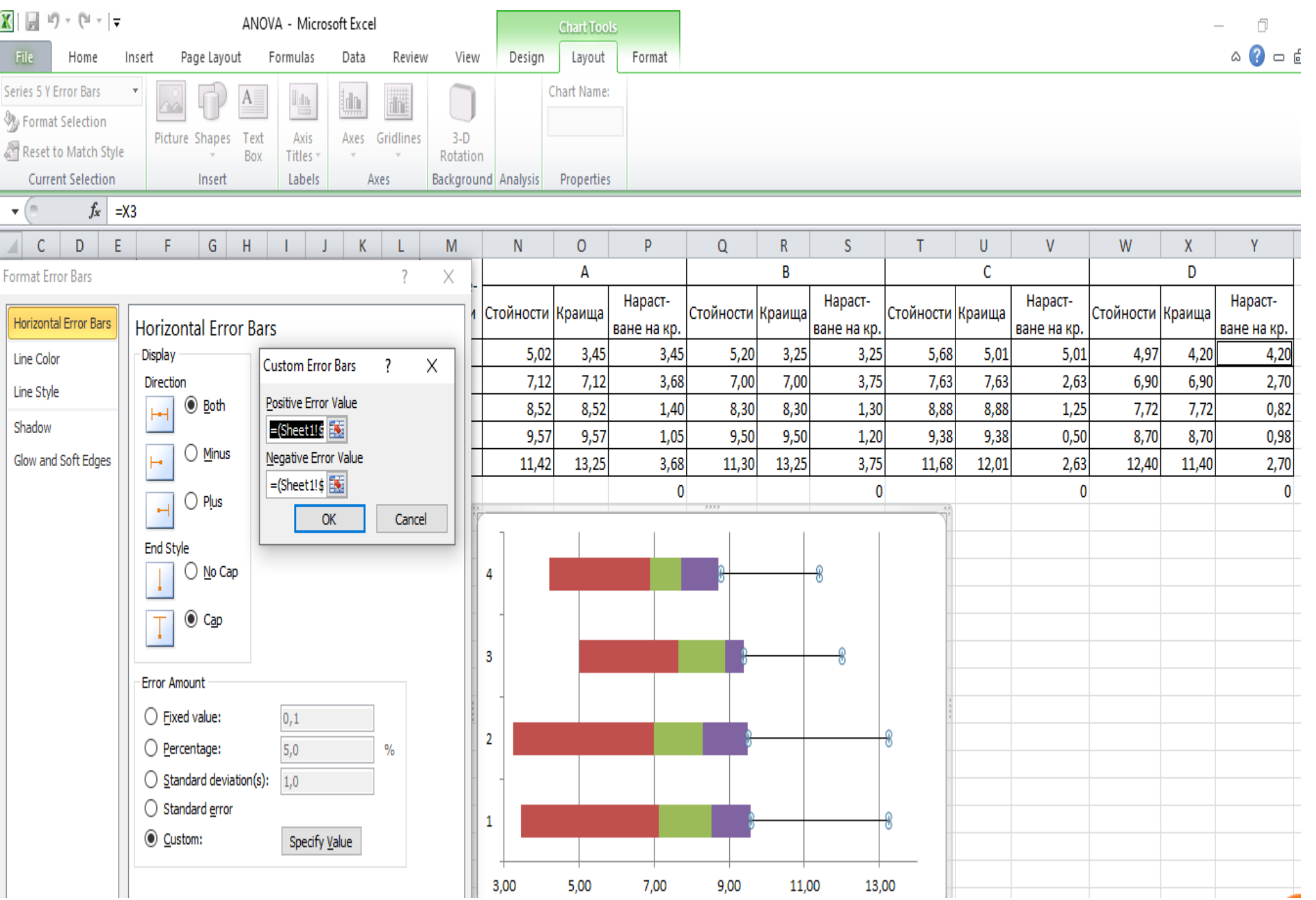
Маркираме с двоен клик върху някой от десните мустачки. После

Layout -> Error bar -> Error bars with standard error

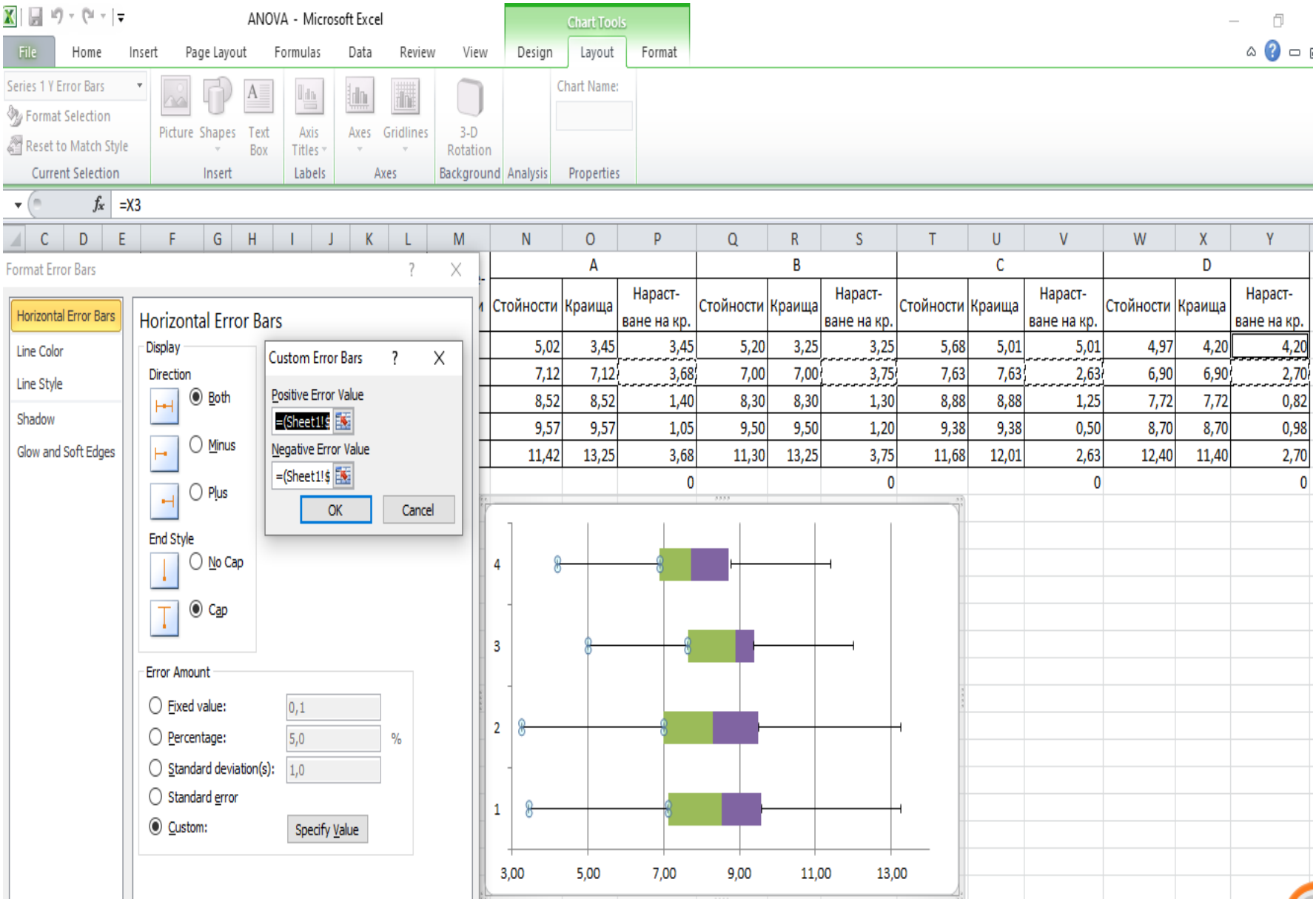
От появилия се прозорец избираме Custom и Specify value

както е показано на следващата фигура.

По-голямо внимание е нужно на клетката Positive Error Value и клетката Negative Error Value.

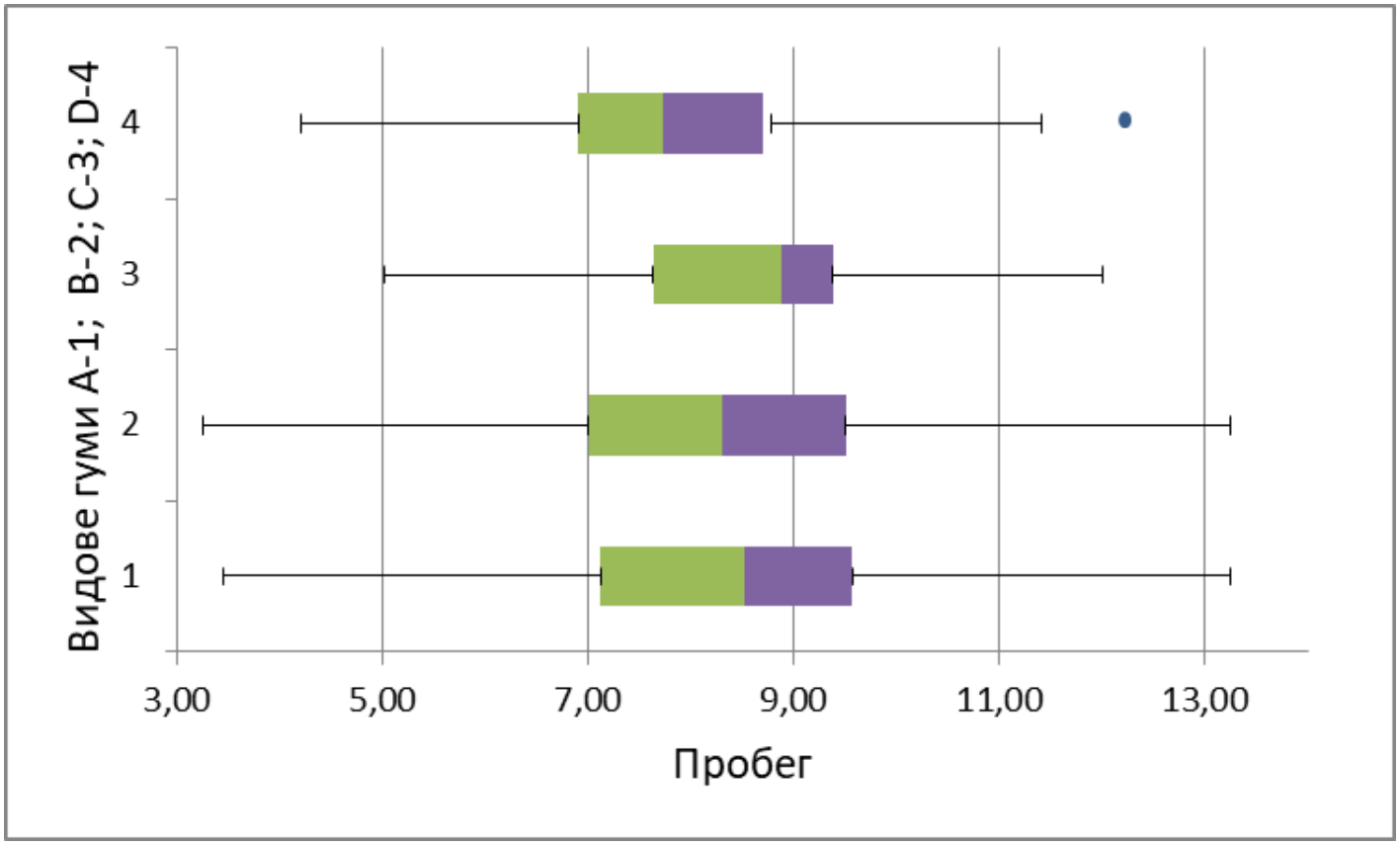


В клетката Positive Error Value маркираме нулите, а в клетката Negative Error Value попълваме чрез кликване върху тях адресите на P7,S7,V7 и Y7.



По аналогичен начин правим левите мустачки. При тях в клетката Negative Error Value маркираме нулите, а в клетката Positive Error Value попълваме чрез кликване върху тях адресите на клетките P4,S4,V4 и Y4.

Получаваме



Вече видяхме, че само при вида гуми D имаме една силно отличаваща се стойност.

Наблюдава се, че четирите типа гуми имат пробег с подобно поведение.

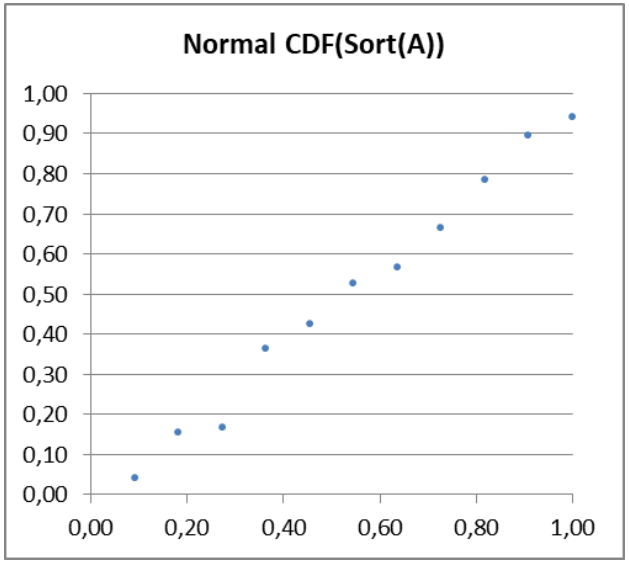
По-точно ще проверим това със средствата на дисперсионния анализ.

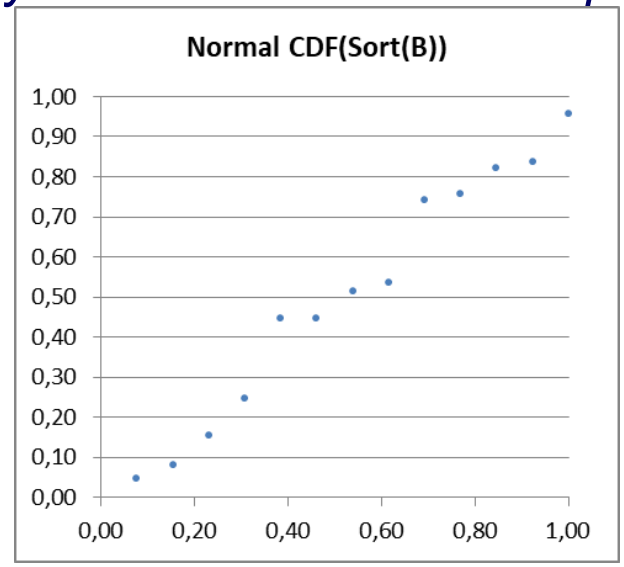
в) Като използвате четири pp-plot проверете дали наблюдаваните четири условни разпределения са нормални.

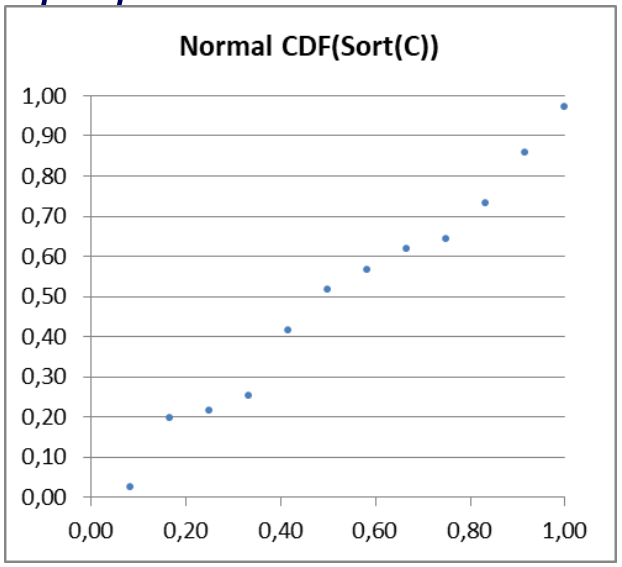
Решение: Във всяка колонка от изходната таблица с данни подреждаме наблюденията възходящо по големина без да разширяваме селекцията. Построяваме и останалите колонки в следващата таблица. Като например клетката Normal CDF(Sort(A)) е получена чрез =NORM.DIST(B2;AVERAGE(B:B);STDEV(B:B);1)

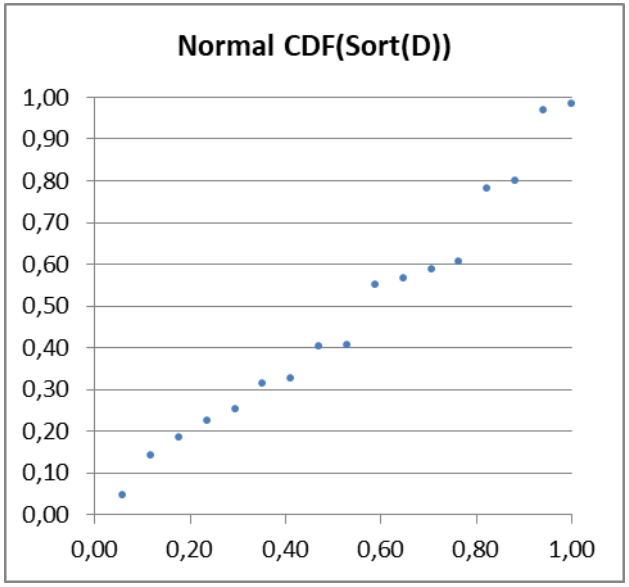
i	Sort(A)	i/n1	Normal CDF(Sort(A))	Sort(B)	i/n2	Normal CDF(Sort(B))	Sort(C)	i/n3	Normal CDF(Sort(C))	Sort(D)	i/n4	Normal CDF(Sort(D))
1	5,02	0,09	0,04	5,20	0,08	0,05	5,68	0,08	0,03	4,97	0,06	0,05
2	6,42	0,18	0,15	5,70	0,15	0,08	7,38	0,17	0,20	6,10	0,12	0,14
3	6,52	0,27	0,17	6,40	0,23	0,15	7,48	0,25	0,21	6,45	0,18	0,19
4	7,72	0,36	0,36	7,00	0,31	0,25	7,68	0,33	0,25	6,72	0,24	0,23
5	8,02	0,45	0,42	8,00	0,38	0,45	8,38	0,42	0,42	6,90	0,29	0,25
6	8,52	0,55	0,53	8,00	0,46	0,45	8,78	0,50	0,52	7,24	0,35	0,31
7	8,72	0,64	0,57	8,30	0,54	0,51	8,98	0,58	0,57	7,31	0,41	0,33
8	9,22	0,73	0,67	8,40	0,62	0,54	9,18	0,67	0,62	7,71	0,47	0,40
9	9,92	0,82	0,78	9,40	0,69	0,74	9,28	0,75	0,64	7,72	0,53	0,41
10	10,82	0,91	0,89	9,50	0,77	0,76	9,68	0,83	0,73	8,41	0,59	0,55
11	11,42	1,00	0,94	9,90	0,85	0,82	10,38	0,92	0,86	8,50	0,65	0,57
12				10,00	0,92	0,84	11,68	1,00	0,97	8,60	0,71	0,59
13				11,30	1,00	0,96				8,70	0,76	0,61
14										9,68	0,82	0,78
15										9,80	0,88	0,80
16										11,73	0,94	0,97
17										12,40	1,00	0,99

Като използваме третата и четвъртата колонка маркираме ги и после Insert -> Scatter получаваме първата от следващите pp-plot. По аналогичен начин са получени и останалите три графики.









Наблюдаваме,че точките от тези pp-plot са разположени около ъглополовящата на първи квадрант. Следователно можем да приемем, че наблюдаваните условни разпределения в четирите групи са нормални.

г) С риск за грешка от първи род $\alpha = 0.05$, проверете хипотезата, че вида на гумите не е статистически значим за пробега им.

Решение: Формулираме нулевата и алтернативната хипотези.

H_0 : $E(\eta|\xi = X_1) = E(\eta|\xi = X_2) = E(\eta|\xi = X_3) = E(\eta|\xi = X_4)$, т.е. отклоненията между средните в различните групи се дължат на случайни, кратко временно действащи фактори, т.е. влиянието на вида на гумите върху техния пробег не е статистически значимо.

Алтернативата е

H_1 : Поне две от средните $E(\eta|\xi = X_1)$, $E(\eta|\xi = X_2)$, $E(\eta|\xi = X_3)$, $E(\eta|\xi = X_4)$, са различни, т.е. отклоненията между поне две от средните в различните групи се дължат на системно действащи фактори, т.е. влиянието на вида на гумите върху техния пробег е статистически значимо.

Избираме риск за грешка от първи род $\alpha = 0.05$.

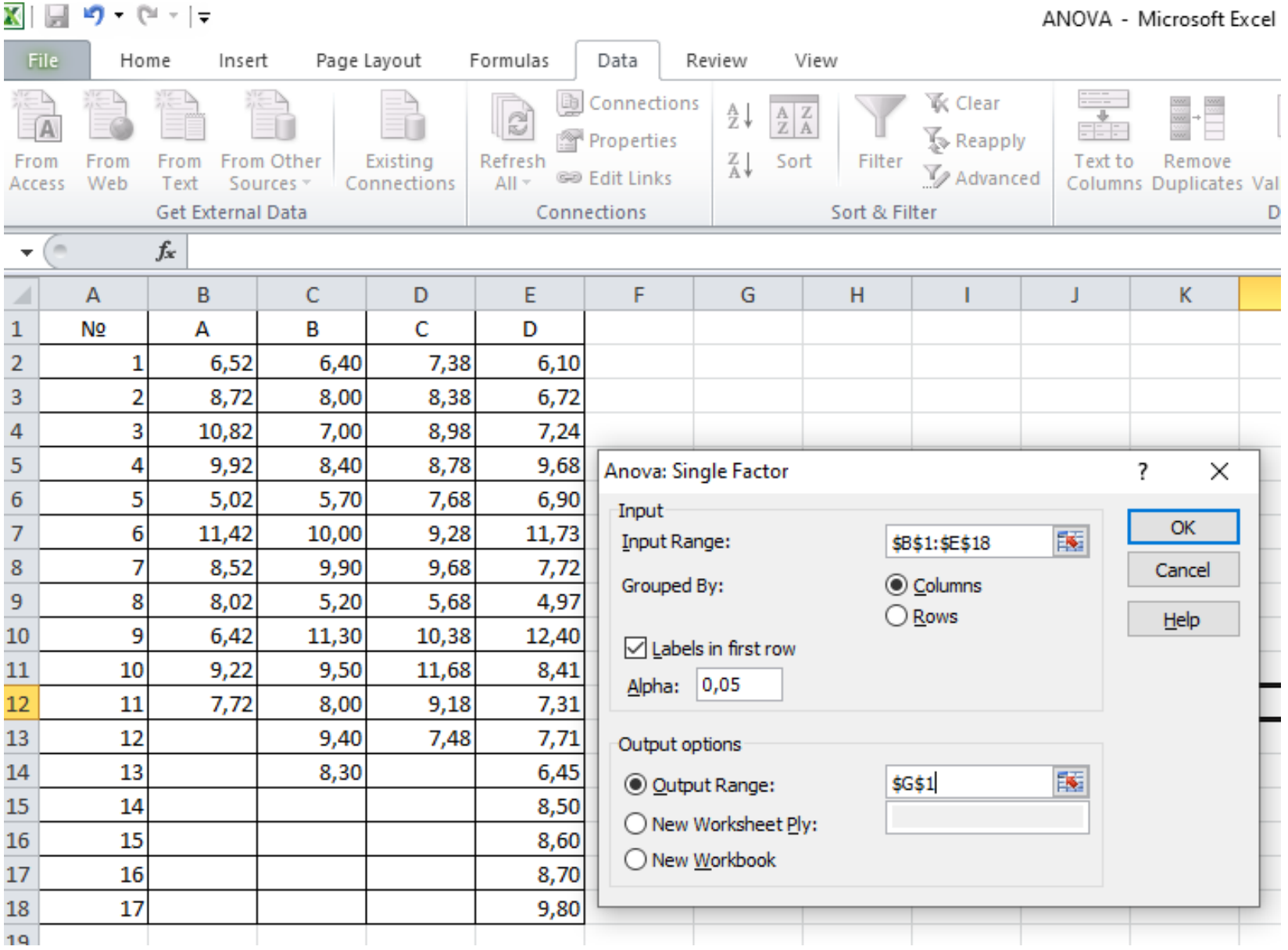
Критичната област в случая има вида

$$W_\alpha = \left\{ (x_1,\ldots,x_n) \in \aleph : \frac{\dot{S}_M^2}{\dot{S}_B^2} \geq x_{1-\alpha,F(k-1,n-k)} \right\} =$$

$$= \left\{ (x_1,\ldots,x_n) \in \aleph : \frac{\dot{S}_B^2}{\dot{S}_M^2} < x_{1-\alpha,F(n-k,k-1)} \right\}.$$

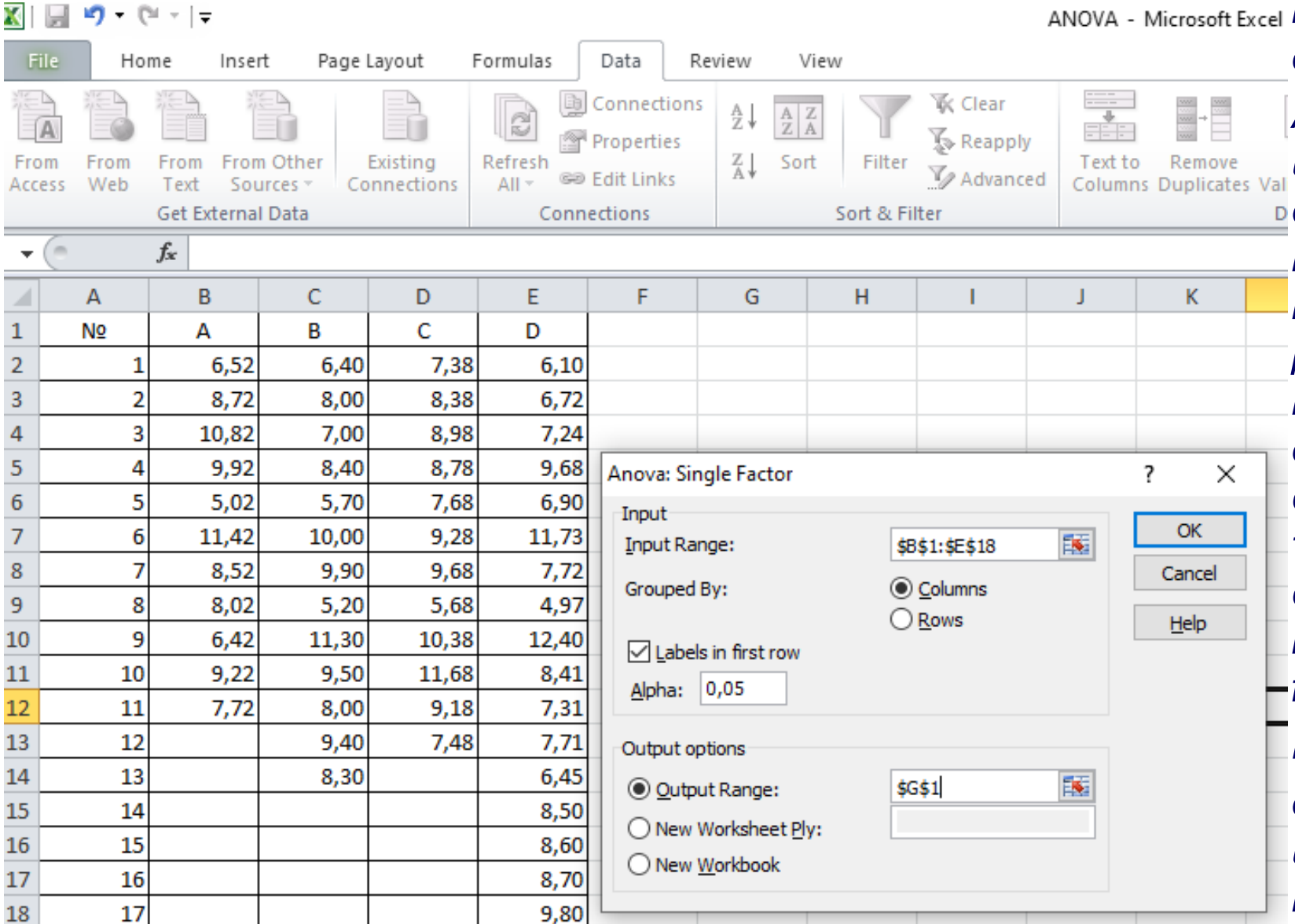
Предполагаме, че данните са въведени както в таблицата по-долу.

За да проверим дали извадката попада в критичната област за H_0 един от начините е да използваме менюто Data - > Data Analysis - > Anova: Single Factor



	A	B	C	D	E
1	№	A	B	C	D
2	1	6,52	6,40	7,38	6,10
3	2	8,72	8,00	8,38	6,72
4	3	10,82	7,00	8,98	7,24
5	4	9,92	8,40	8,78	9,68
6	5	5,02	5,70	7,68	6,90
7	6	11,42	10,00	9,28	11,73
8	7	8,52	9,90	9,68	7,72
9	8	8,02	5,20	5,68	4,97
10	9	6,42	11,30	10,38	12,40
11	10	9,22	9,50	11,68	8,41
12	11	7,72	8,00	9,18	7,31
13	12		9,40	7,48	7,71
14	13		8,30		6,45
15	14				8,50
16	15				8,60
17	16				8,70
18	17				9,80

След натискането на OK се появява диалоговият прозорец, показан по-долу. В полето Input Range попълваме чрез маркиране от горна лява до долна дясна клетка правоъгълната област от клетки, заети от изходните данни.



	A	B	C	D	E
1	№	A	B	C	D
2	1	6,52	6,40	7,38	6,10
3	2	8,72	8,00	8,38	6,72
4	3	10,82	7,00	8,98	7,24
5	4	9,92	8,40	8,78	9,68
6	5	5,02	5,70	7,68	6,90
7	6	11,42	10,00	9,28	11,73
8	7	8,52	9,90	9,68	7,72
9	8	8,02	5,20	5,68	4,97
10	9	6,42	11,30	10,38	12,40
11	10	9,22	9,50	11,68	8,41
12	11	7,72	8,00	9,18	7,31
13	12		9,40	7,48	7,71
14	13		8,30		6,45
15	14				8,50
16	15				8,60
17	16				8,70
18	17				9,80

Ако имаме етикети на първия ред, както е в случая слагаме отгълче срещу Labels in first row.

Въвеждаме избраното α.

В полето Output Range избираме клетка, която ще съдържа горния ляв ъгъл на таблиците с резултата. Натискаме OK и получаваме:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	11	92,32	8,392727	3,756182
B	13	107,1	8,238462	3,242564
C	12	104,56	8,713333	2,43697
D	17	138,9373	8,172783	3,708428

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2,297302	3	0,765767	0,230746	0,874534	2,793949
Within Groups	162,6141	49	3,318655			
Total	164,9114	52				

Колоните от първата таблица съдържат съответно:
Count – броевете на наблюденията в отделните подгрупи;
Sum - сумата от наблюденията в отделните подгрупи;
Average – средните аритметични в отделните подгрупи;
Variance – дисперсиите в отделните подгрупи.
Наблюдаваме приблизително еднакви средни и еднакви дисперсии в подгрупите. Т.е. можем да очакваме след малко приемане на H_0 .

Втората таблица съдържа съответно следните стойности:

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2,2973024	3	0,765767475	0,230746321	0,874534	2,793948852
Within Groups	162,6141	49	3,318655187			
Total	164,91141	52				

Източник на дисперсията	*Девиация*	*Степени на свобода*	*Дисперсия*	*Емпирична х-ка*	*P-value*	$x_{1-\alpha,F(k-1,n-k)}$
Междугрупова	SS_M	k-1	$\dot{S}_M^2 = \frac{SS_M}{k-1}$	$\frac{\dot{S}_M^2}{\dot{S}_B^2}$		
Вътрешнорупова	SS_B	n-k	$\dot{S}_B^2 = \frac{SS_B}{n-k}$			
Обща	SS_0	n-1	$\dot{S}_0^2 = \frac{SS_0}{n-1}$			

Тъй като емпиричната характеристика е по-малка от теоретичната характеристика $x_{1-\alpha,F(k-1,n-k)}$ *, то извадката не е в критичната област за* H_0 *за това нямаме основание да отхвърлим* H_0*. Т.е. вида на гумите не влияние статистически значимо на техния пробег.*

Същият извод можем да направим и с помощта на p-value.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2,2973024	3	0,765767475	0,230746321	0,874534	2,793948852
Within Groups	162,6141	49	3,318655187			
Total	164,91141	52				

Източник на дисперсията	*Девиация*	*Степени на свобода*	*Дисперсия*	*Емпирична х-ка*	*P-value*	$x_{1-\alpha,F(k-1,n-k)}$
Междугрупова	SS_M	k-1	$\dot{S}_M^2 = \frac{SS_M}{k-1}$	$\frac{\dot{S}_M^2}{\dot{S}_B^2}$		
Вътрешнорупова	SS_B	n-k	$\dot{S}_B^2 = \frac{SS_B}{n-k}$			
Обща	SS_0	n-1	$\dot{S}_0^2 = \frac{SS_0}{n-1}$			

Тъй като стойността на полученото p-value е по-голяма от α = 0.05, то извадката не е в критичната област за H_0 за това нямаме основание да отхвърлим H_0. Т.е. вида на гумите не влияние статистически значимо на техния пробег.

22. Корелационен анализ. Основни понятия.

След усвояването на информацията от тази лекция Вие ще знаете

- ✓ *Какво е корелационна зависимост?*
- ✓ *Каква е разликата между корелационна и функционална зависимост?*
- ✓ *Как да определяте силата на зависимостта между два признака?*

1. Основни понятия. При изследване на зависимости между статистически признаци обикновено се решават две задачи. Едната е определяне на ***формата на зависимостта****, а другата е определяне на* ***силата на зависимостта****. Първата е обект на* ***Регресионния анализ****, с който ще се запознаем в следващата тема, а втората е обект на* ***Корелационния анализ****.*

Детерминистичната математика се занимава основно с изучаването на ***“функционални зависимости”,*** *т.е. на един фиксиран аргумент на функцията се съпоставя винаги едно и също детерминирано, множество от числа, най-често точно едно число. В тази тема ще се научим да измерваме корелационна зависимост или връзка. Това е зависимост, при която на едно фиксирано значение на единия признак, съответства множество от значения на другия, всяко от които с определена вероятност. С корелационния коефициент се измерва силата на връзката или зависимостта между разглежданите признаци. Чрез него можем да отговорим на въпроса:*

- *До каква степен като изменяме едната величина се изменя другата?*

Възможно е обаче тази зависимост да не е причинно-следствена и да се предизвиква или влияе и от други неизследвани признаци, ето защо трябва да бъдем особено внимателни при анализиране на резултатите.

В зависимост от броя на променливите, между които ще мерим силата на корелационната зависимост, говорим за еднофакторен или многофакторен корелационен анализ.

Ние ще разгледаме само еднофакторен корелационен анализ.

Определянето на методологията за пресмятане на корелационния коефициент става в зависимост от вида на скалата, по която са отчетени значенията на изследвания признак.

Ако и двата признака са метрирани можем да пресметнем корелационния коефициент на Браве. Ако двата признака са представени на рангова скала може да се използват коефициентите на корелация на Спирмън или Кендал. В общия случай можем да използваме коефициентите на взаимносвързаност (на контингенция) на Пирсън и Чупров и като техен частен случай при две дихотомни скали се използва коефициента на четириклетъчна корелация на Пирсън. При един дихотомен и един метриран признак са подходящи бисериалните коефициенти на корелация и т.н. До част от тези резултати учените са достигнали по емпиричен път.

Ако поне единият признак е метриран той започва обикновено с изчертаване на подходящи графики за онагледяване на зависимостите между наблюдаваните признаци.

- Ако единият е метриран а другият категориен, е удачно да се наблюдават паралелни графики с мустачки на разпределението на метрирания признак в групите, оформени от различните измерени значения на неметрирания признак.

- Ако и двата признака са метирани, първо се наблюдава корелационното поле или т.нар. двумерно разпределение на данните.

От графичния образ на тези графики получаваме първична представа за очакваните резултати.

2. Измерване на зависимости при неинтервални скали

а) Рангови коефициенти на корелация.

Да предположим, че над единиците от съвкупността са извършени наблюдения, върху два признака, измерени на рангова скала.

- Рангов коефициент на корелация на Спирмън.

За да използваме този коефициент, ранговете по един и същ признак трябва да са различни числа от 1 до n. Спирмън използва като измерител на близостта на ранговете, сумата от квадратите на разликите им d_i, i = 1, 2, …, n. Ако съществува силна положителна зависимост между ранговете на единиците, те би трябвало да съвпадат и сумата от квадратите на разликите им би била нула. Ако зависимостта е силна отрицателна, ранговете ще са подредени в обратен ред. Разликите им в този случай, ако n е четно, ще образуват редица само от нечетните числа от – (n-1) до (n-1) и при n = 2k сумата от квадратите им ще е

$$\sum_{i=1}^{n} d_i^2 = 2\sum_{i=1}^{k} d_i^2 = 2\sum_{i=1}^{k} (2i-1)^2 = \frac{n(n^2-1)}{3}.$$

Ако n е нечетно разликите на ранговете ще образуват редица само от четните числа в същия интервал и при n = 2k-1

$$\sum_{i=1}^{n} d_i^2 = 2\sum_{i=1}^{k-1} d_i^2 = 2\sum_{i=1}^{k-1} (2i)^2 = \frac{n(n^2-1)}{3}.$$

При липсата на каквато и да е зависимост можем да приемем, че тази сума ще е средното аритметично на двете крайни възможности, т.е.

$$\frac{0+\frac{n(n^2-1)}{3}}{2}=\frac{n(n^2-1)}{6}.$$

Като отнесем тази величина към действителната сума от квадратите на разликите получаваме измерител на зависимостта, който обаче би бил нула при силна правопропорционална зависимост между ранговете. Ето защо ранговият коефициент на корелация на Спирмън се пресмята по формулата

$$r_{Cn}=1-\frac{\sum_{i=1}^{n}d_i^2}{\frac{n(n^2-1)}{6}}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}.$$

- Коефициент на корелация на Кендал, τ. За да разкажем как се пресмята този коефициент, трябва да дефинираме понятията съответствия и инверсии. Да предположим, че статистическите единици са подредени по ранговете на признака X във възходящ ред. Брой на съответствията p_i на i – тата статистическа единица, се нарича броят на двойките след i – тата, т.е. за j = i+1,…,n такива че $X_i < X_j$ и $Y_i < Y_j$. Брой на инверсиите q_i на i – тата статистическа единица, се нарича броят на двойките след i – тата, т.е. при j = i+1,…, n, за които $X_i < X_j$ и $Y_i > Y_j$. Ако всички двойки са разположени в еднакъв порядък, възходящо, сумата от всички съответствия P ще е равна на сумата на естествените числа от 1 до n-1, т.е.

$$P = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n-1} i = \frac{n(n-1)}{2}.$$

По аналогичен начин ако ранговете на признака Y са подредени низходящо, сумата от инверсиите Q ще е

$$Q = \sum_{i=1}^{n} q_i = \frac{n(n-1)}{2}.$$

Като измерител на зависимостта между двата статистически признака Кендал използва отношението на разликата между съответствията и инверсиите и сумата на естествените числа от 1 до n-1, т.е. неговият корелационен коефициент има вида

$$\tau = \frac{2(P-Q)}{n(n-1)}.$$

Преимуществото му пред коефициента на Спирмън е, че може да се използва и при еднакви рангове.

Двата коефициента, на Спирмън и Кендал, в общия случай имат различни стойности.

б) Коефициенти на взаимосвързаност (на контингенция) на Пирсън, Крамер и Чупров.

Да предположим, че признаците X и Y имат значения съответно $X_1,\ldots,X_k$ и $Y_1,\ldots,Y_s$, и че резултатите от групировката след наблюдението върху тях са разположени в следната таблица:

Y / X	Y_1	...	Y_s	*Общо:*
X_1	f_{11}	...	f_{1s}	f^{X}_{1}
...	...	...	...	...
X_k	f_{k1}	...	f_{ks}	f^{X}_{k}
Общо:	f^{Y}_{1}	...	f^{Y}_{s}	n

където с f_{ij} е означен броят на статистическите единици притежаващи i – тото значение на признака X и j – тото значение на признака Y. f^{X}_{1} е честотата в първата група на признака X. По аналогичен начин до f^{X}_{k}. f^{Y}_{1} е честотата в първата група на признака Y и т.н. до f^{Y}_{s}. n е броят на всички наблюдавани единици.

Тогава в сила са следните съотношения

$$n = \sum_{i=1}^{k}\sum_{j=1}^{s} f_{ij}, \qquad f_i^{\,X} = \sum_{j=1}^{s} f_{ij}, \qquad f_j^{\,Y} = \sum_{i=1}^{k} f_{ij}.$$

Преди да преминем към изчисляването на двата коефициента, трябва да определим теоретичните честоти в групите $\hat{f}_{ij} = \dfrac{f_i^{\,X} f_j^{\,Y}}{n}.$

След това определяме χ^2 *характеристиката и* φ^2 *по формулите*

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{s} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}, \quad \varphi^2 = \frac{\chi^2}{n}.$$

Коефициентът на Пирсън *се пресмята по формулата*

$$r_{П} = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}.$$

Този коефициент е равен на 0 ако двата наблюдавани признака не са свързани. Недостатъкът му е, че максималната му стойност е

$$r_{П,\max} = \sqrt{\frac{m}{1+m}},$$

където m = min(k, s)-1.

Коефициентът на Крамер *се пресмята по формулата*

$$r_{Кр.} = \sqrt{\frac{\chi^2}{nm}} = \sqrt{\frac{\varphi^2}{m}},$$

където m = min(k, s)-1.

Преимуществото му пред този на Пирсън е, че коефициент на Крамер е равен на 0 ако двата наблюдавани признака не са свързани и на 1, ако те са свързани.

Коефициентът на Чупров *е нормиран със степените на свобода на* χ^2 *характеристиката. Изчислява се от*

$$r_Ч = \sqrt{\frac{\varphi^2}{(k-1)(s-1)}}.$$

При несвързани признаци е равен на 0. Както коефициентът на Пирсън, така и тук ако двата наблюдавани признака са свързани

$$r_{Ч,\max} = \sqrt{\frac{m}{(k-1)(s-1)}}.$$

Да обърнем внимание на това, че тези коефициенти са само положителни. Това не води до недостатъци при анализа им, защото те се използват основно при работа със слаби скали, т.е. при определяне на силата на зависимостта при неметрирани признаци. В този случай не е удачно да говорим за права или обратна пропорционалност.

***В случая на два дихотомни признака**, т.е. ако k = 2 и s = 2 корелационната таблица има вида*

Y / X	Y_1	Y_2	*Общо:*
X_1	a	b	$a+b$
X_2	c	d	$c+d$
Общо:	$a+c$	$b+d$	n

и коефициентът на Пирсън се нарича четириклетъчен и се свежда до

$$r_{П} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

3. Коефициент на праволинейна корелация на Браве

Да разгледаме двумерна, проста извадка $(X_1, Y_1), \ldots, (X_n, Y_n)$, *т.е. при всички статистически единици се наблюдават значенията на два признака, т.е. правят се наблюдения върху две случайни величини* ξ *и* η.

Наблюденията при отделните статистически единици са независими.

Коефициентът на праволинейна корелация на Браве *се определя по формулата*

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2}}.$$

Както си личи от названието му, този коефициент измерва до колко точките от корелационното поле се групират около права. В случая на репрезентативна извадка, корелационният коефициент е точкова оценка за $\mathrm{cor}(\xi, \eta)$.

Когато данните са групирани, т.е. представени в корелационна таблица се прилага следната формула:

$$\dot{r} = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \dot{x}_i \dot{y}_j f_{ij} - n\overline{X}_n \overline{Y}_n}{nS_X S_Y},$$

където k_1, k_2 са броевете на групите оформени съответно по признаците ξ и η. $\dot{x}_1, \dot{x}_2, \ldots, \dot{x}_{k_1}$ са класовите представители на признака ξ. $\dot{y}_1, \dot{y}_2, \ldots, \dot{y}_{k_2}$ са класовите представители на признака η. S_X и S_Y са съответните стандартни отклонения.

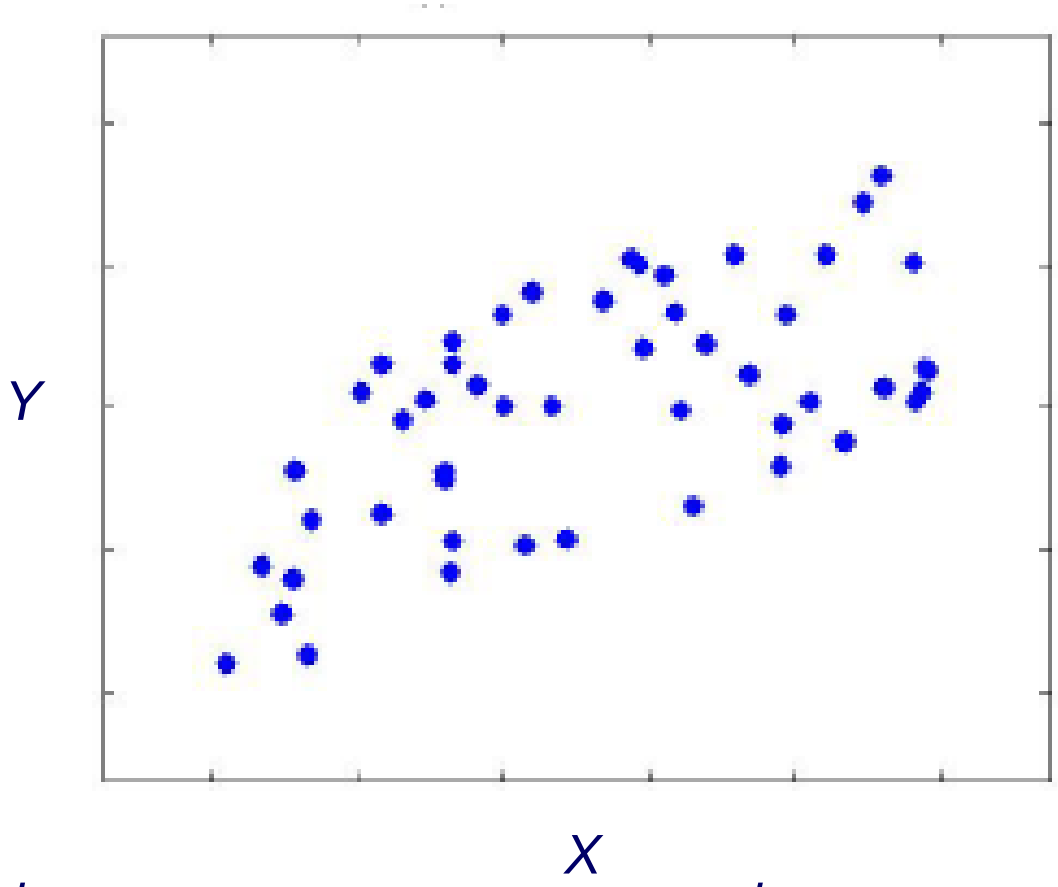
С f_{ij} е означен броя на статистическите единици, попаднали в i – тата група на признака ξ и в j – тата група на признака η.

$$n = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} f_{ij}.$$

и представлява броя на всички наблюдавани единици.

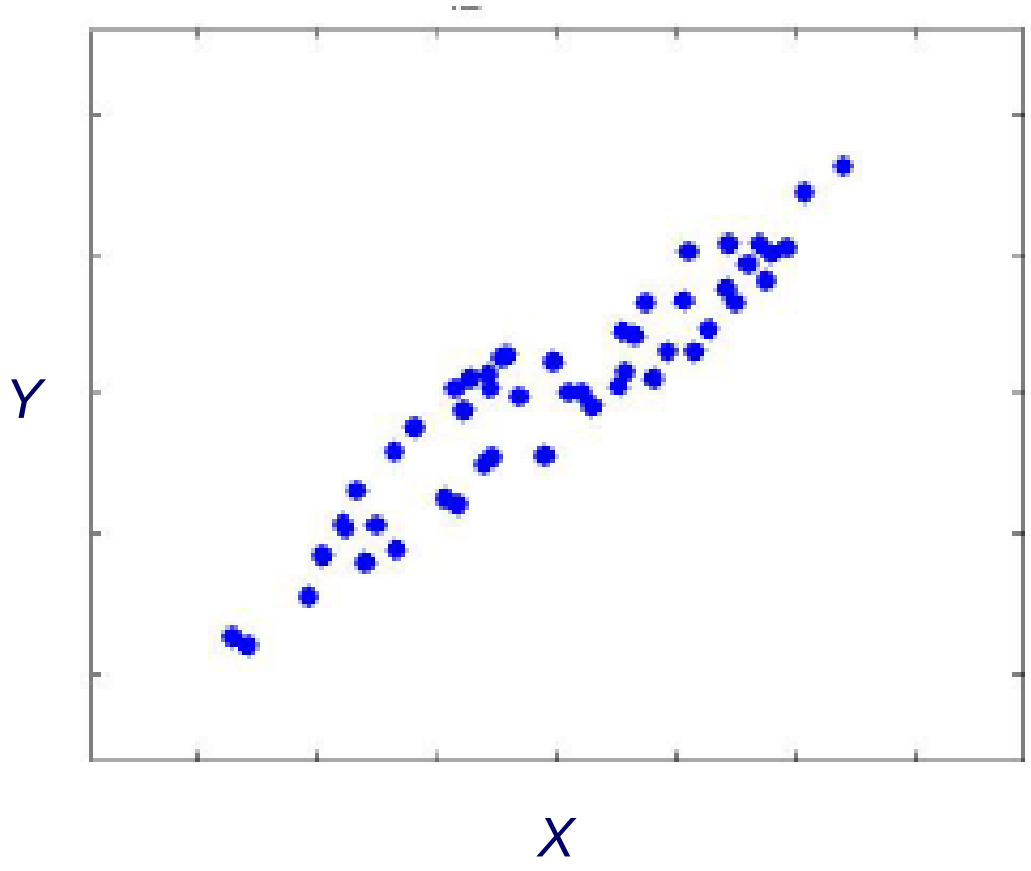
Този корелационен коефициент е винаги в интервала [-1,1].

Ако той е положителен имаме правопропорционална зависимост между значенията на двата признака, и точките от корелационното поле се групират около възходяща права.



На тази фигура корелационният коефициент на Браве е 0,6117.

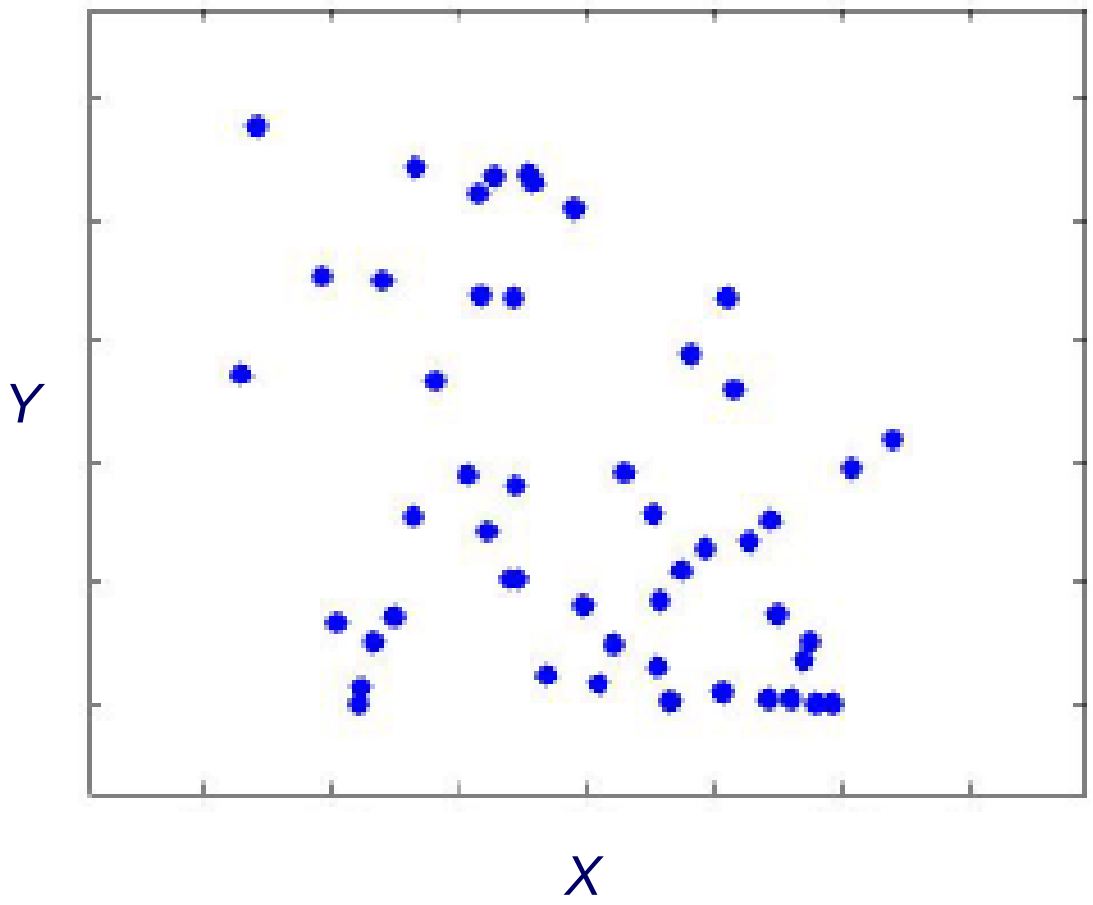
Колкото корелационният коефициент на Браве е по-близо до +1, толкова точките от двумерното разпределение на данните са по-силно концентрирани върху правата.



На фигурата в ляво той е 0,9386.

Ако корелационният коефициент на Браве е точно 1, между наблюдаваните признаци има функционална зависимост. Нещо повече, съществуват коефициенти $a \in R$ *и* $b > 0$, *такива, че* $P(\xi = a + b\eta) = 1$.

Ако r < 0 зависимостта е обратнопропорционална и правата около, която се групират точките е низходяща.

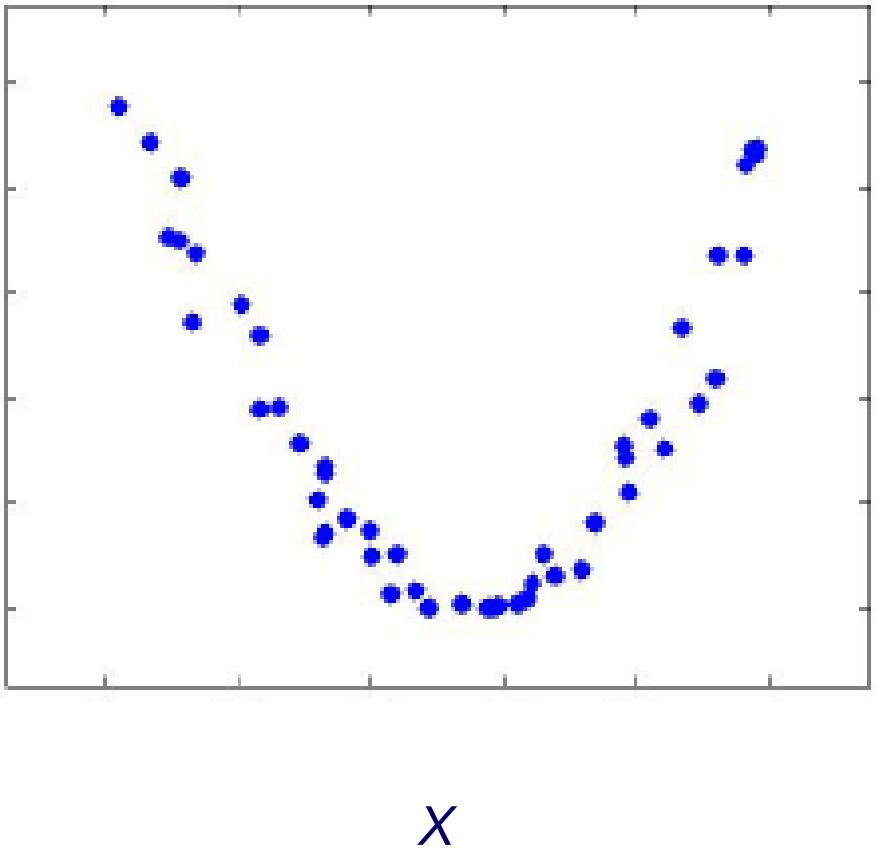


На графиката в ляво r = -0,3680.

Когато r = -1, между наблюдаваните признаци има функционална, праволинейна, обратнопропорционална зависимост.

Да припомним, че това изобщо не значи, че тя е причинно- следствена.

При нормално разпределени съвкупности r = 0 е еквивалентно на независимост на признаците. В общия случай при анализирането на нулев корелационен коефициент трябва да бъдем особено внимателни.



Ако измерваните величини са независими, то безусловно r = 0, обратното обаче не винаги е вярно. Ето защо, когато r = 0 можем да кажем само, че между X и Y не може да съществува праволинейна зависимост, но криволинейна може. В този случай говорим само за некорелираност на разглежданите признаци. На тази фигура r = -0,0319.

Квадратът r^2 на корелационния коефициент r се нарича ***коефициент на детерминация****. Той показва каква част, или ако е превърнат в проценти, колко процента от вариацията на единия, зависимия признак се дължи на вариацията на другия, независимия признак.*

1 – r^2 се нарича ***коефициент на индетерминация****, неопределеност (ако работим в проценти умножаваме по 100). Показва каква част от вариацията на зависимата променлива се дължи на други, неразглеждани в модела фактори.*

Корелационният коеф. на Браве по негрупирани данни се смята например с помощта на функцията CORREL(Масив1; Масив2) в Excel.

23. Регресионен анализ

След усвояването на информацията от тази лекция Вие ще можете:

- ✓ *Да моделирате формата на влиянието на един независим метриран, факторпризнак X върху друг зависим също метриран, резултативен признак Y на единиците от съвкупността;*
- ✓ *Да оценявате значенията на резултативния признак Y като знаете значението на факторпризнака X;*
- ✓ *Да определяте изменението на резултативния признак Y като знаете изменението на факторпризнака X.*

Със средствата на регресионния анализ се моделира формата на зависимостта на една зависима, резултативна величина Y от един или няколко факторпризнаци, като не се отчита, че изменението на разглежданите величини може да се дължи на външни, невключени в модела признаци. Ако факторпризнакът е един, говорим за единична регресия. Иначе говорим за множествена регресия. Тук ще се спрем на методологията на единичната регресия.

1. Същност на регресионния анализ

Еднофакторният регресионен анализ обикновено започва с изчертаване на корелационно поле (емпиричното двумерно разпределение) на данните. По абсцисната ос се нанасят значенията на фактор-признака X, а по ординатната, тези на резултативния признак Y. По графичния образ на това поле избираме линия, която най-добре ще приближава точките му, т.е. такава, с която ще моделираме регресията. Трябва да знаем аналитичното й представяне. Нека то да бъде

(1) $$\hat{y} = \varphi(x, \mathbf{a}),$$

където координатите на вектора $\mathbf{a} = (a_1, a_2, \ldots, a_d)$ *са неизвестните параметри за функцията* φ. *Това уравнение се нарича* ***уравнение на линията на регресия***. *Уравнението*

(2) $$y = \varphi(x, \mathbf{a}) + \varepsilon,$$

където ε *е стохастичната грешка на модела, т.е. тази, която приемаме, че се дължи на случайния характер на извадката и е такава, че математическото й очакване* $E\varepsilon = 0$, $D\varepsilon = \sigma^2$ *и* ε *не зависи от фактор признака X се нарича* ***регресионен модел с адитивна грешка***.

По данните от извадката, обикновено използвайки метода на най-малките квадрати (МНМКв), правим оценка на вектора ***а****. Ще я означаваме с* $\hat{a}$*. Тя минимизира сумата от квадратите на* ***отклонения-та (грешките)*** ε_i ***: =*** $Y_i - \varphi(X_i, a)$*. Т.е.*

$$\underset{a_1, \ldots, a_d}{argmin} \sum_{i=1}^{n} \varepsilon_i^2$$

$$\underset{a_1, \ldots, a_d}{argmin} \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2$$

Тази оценка $\hat{a}$ *се намира като решим относно* $a_1, a_2, \ldots, a_d$, *системата*

$$\left| \begin{array}{c} \dfrac{\partial \sum_{i=1}^{n} (Y_i - \varphi(X_i, a_1, \ldots, a_d))^2}{\partial a_i} = 0 \\ i = 1, \ \ldots, d. \end{array} \right.$$

наречена ***система нормални уравнения***.

След като се определят оценките на параметрите в избрания модел може да се прави проверка на хипотезата за статистическата им значимост.

След това от полученото уравнение на регресия пресмятаме оценките на стойностите на зависимата променлива. Тези оценки ще означаваме с $\hat{y}$, *т.е.* $\hat{y} = \varphi(X, \hat{a}).$

Добре е да тестваме повече от една функция φ. При всяка от тях, при фиксиран факторпризнак X ще получаваме различни оценки на резултативната величина Y. Един от критериите за избор на най-добър модел за съответните данни е да изберем линията, за която сумата от квадратите на отклоненията на фактическите (измерените значения на резултативната величина Y) от техните оценки $\hat{y}$ е минимална. Това е все едно да кажем, че моделът с най-малка ***обща стандартна грешка***

(3)

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - Y_i)^2}{n - d}}$$

е най-добър.

Трябва да отбележим, че освен тази грешка се взема в предвид и разпределението на ε. То трябва да е такова, че да получим добър доверителен интервал за резултативната величина.

След избора на модел се прави проверка на хипотезата, че отклоненията на фактическите стойности от техните оценки имат случаен характер. Проверява се едновременно дали тези остатъци са еднакво разпределени, дали са независими, дали имат симетрично разпределение.

Ако използваме модел с нормално разпределена грешка, както в нашия курс, с някои от критериите за съгласие се проверява дали разпределението на ε е нормално. Проверява се хипотезата за липса на корелация между X и остатъчния компонент ε.

След намирането на уравнението на регресия можем да получим най-добра оценка за Y по зададено значение на X. Това е $\hat{y}$.

В следващите точки на тази тема ще разгледаме по-подробно случаите, когато точките от корелационното поле се групират около права или част от крива от втора степен. В останалите случаи се работи по аналогичен начин.

2. Единична линейна регресия *Да предположим, че изследваме влиянието на фактора X върху резултативния признак Y и, че разполагаме с n на брой двойки от наблюдения*

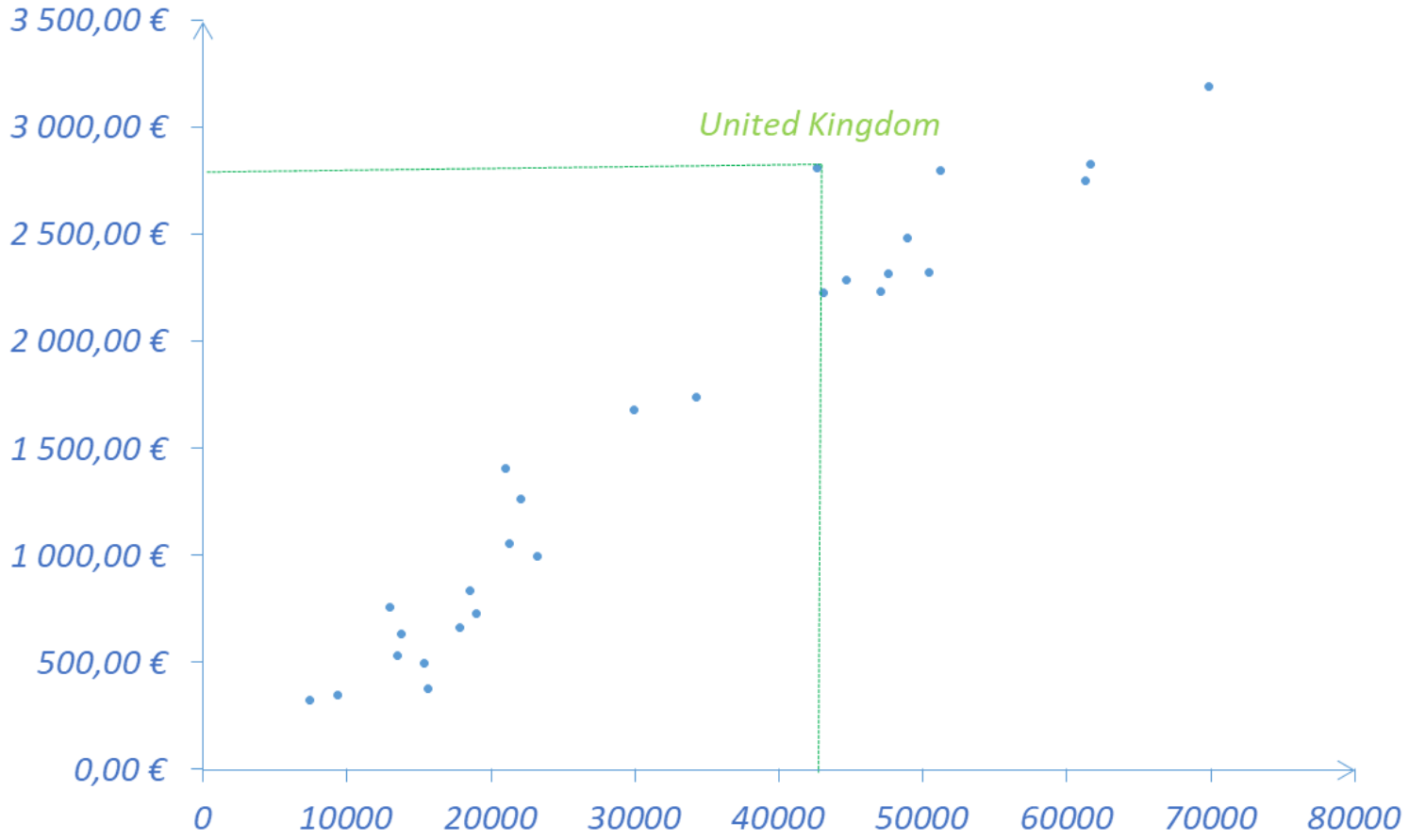
$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

	Страна	БНД	Разполагаем среден месечен доход	Година на присъединяване
1.	Austria	50 390 $	2 320.24 €	1995 г.
2.	Belgium	47 030 $	2 232.85 €	1958 г.
3.	Bulgaria	7 420 $	324.89 €	2007 г.
4.	Croatia	13 020 $	755.21 €	2013 г.
5.	Cyprus			2004 г.
6.	Czech Republic	18 970 $	725.92 €	2004 г.
7.	Denmark	61 310 $	2 751.71 €	1973 г.
8.	Estonia	18 530 $	832.57 €	2004 г.
9.	Finland	48 910 $	2 479.56 €	1995 г.
10.	France	43 070 $	2 223.89 €	1958 г.
11.	Germany	47 640 $	2 315.20 €	1958 г.
12.	Greece	22 090 $	1 262.05 €	1981 г.
13.	Hungary	13 470 $	532.08 €	2004 г.

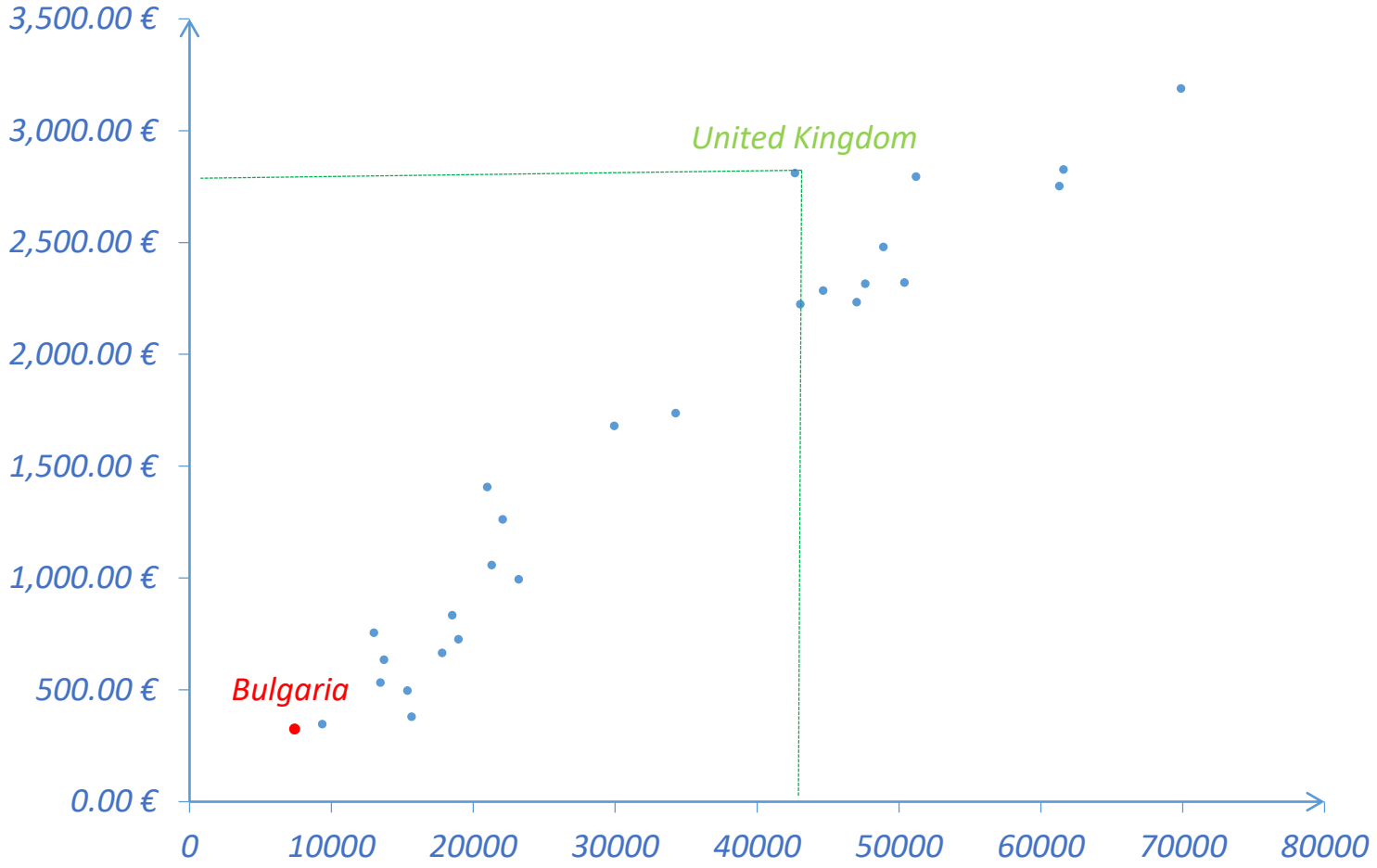
	Страна	БНД	Разполагаем среден месечен доход	Година на присъединяване
14.	Ireland	44 660 $	2 284.40 €	1973 г.
15.	Italy	34 280 $	1 736.18 €	1958 г.
16.	Latvia	15 660 $	379.22 €	2004 г.
17.	Lithuania	15 380 $	496.23 €	2004 г.
18.	Luxembourg	69 880 $	3 187.82 €	1958 г.
19.	Malta	21 000 $	1 406.02 €	2004 г.
20.	Netherlands	51 210 $	2 793.73 €	1958 г.
21.	Poland	13 730 $	634.49 €	2004 г.
22.	Portugal	21 320 $	1 056.92 €	1986 г.
23.	Romania	9 370 $	345.56 €	2007 г.
24.	Slovakia	17 810 $	664.77 €	2004 г.
25.	Slovenia	23 220 $	993.80 €	2004 г.
26.	Spain	29 940 $	1 679.21 €	1986 г.
27.	Sweden	61 600 $	2 825.57 €	1995 г.
28.	United-Kingdom	42 690 $	2 810.26 €	1973 г.

2014 г. https://en.wikipedia.org/wiki/List_of_European_countries_by_average_wage
https://en.wikipedia.org/wiki/List_of_sovereign_states_in_Europe_by_GNI_(nominal)_per_capita

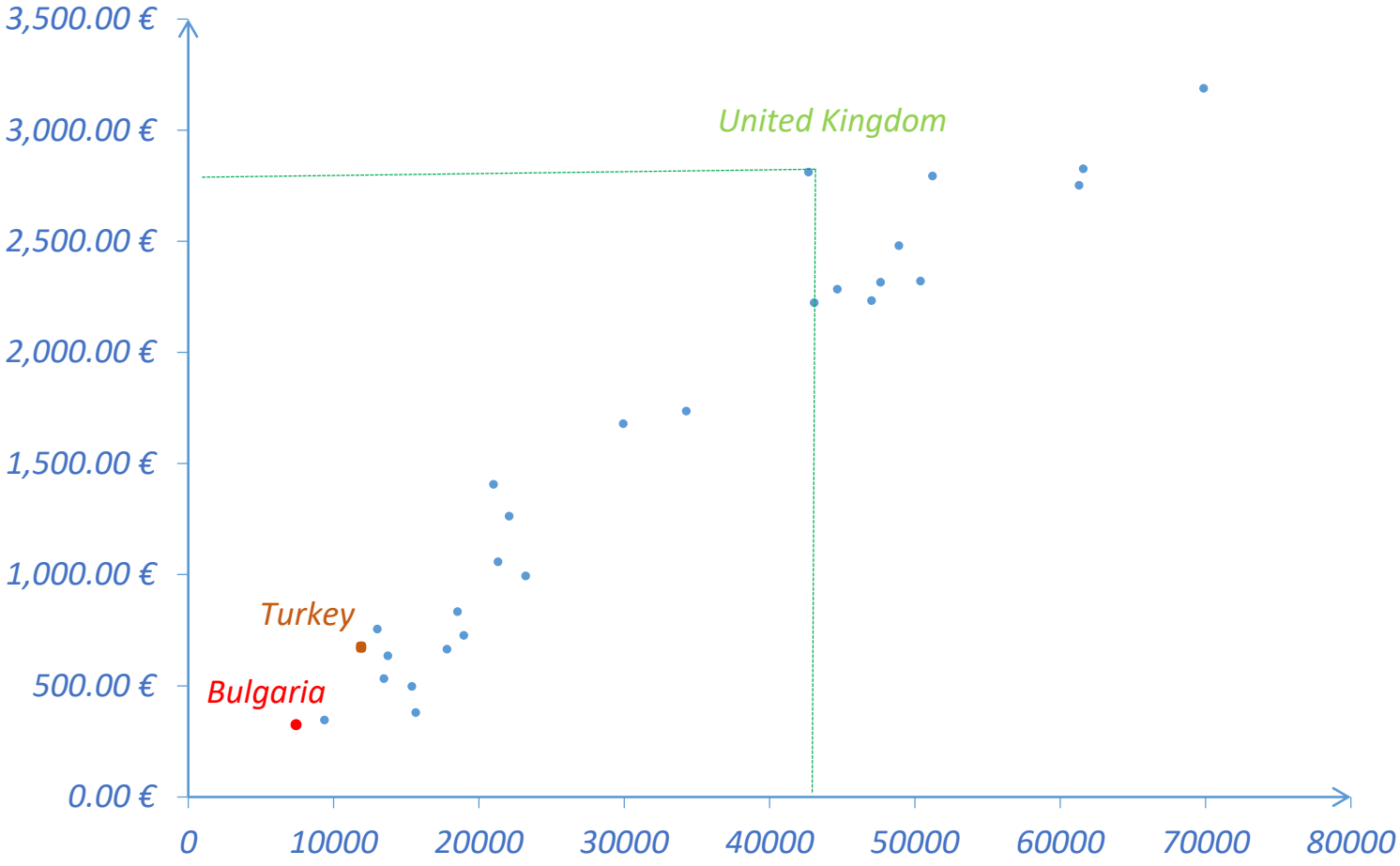
С помощта на Insert ->Scatter първо изчертаваме корелационното поле на данните



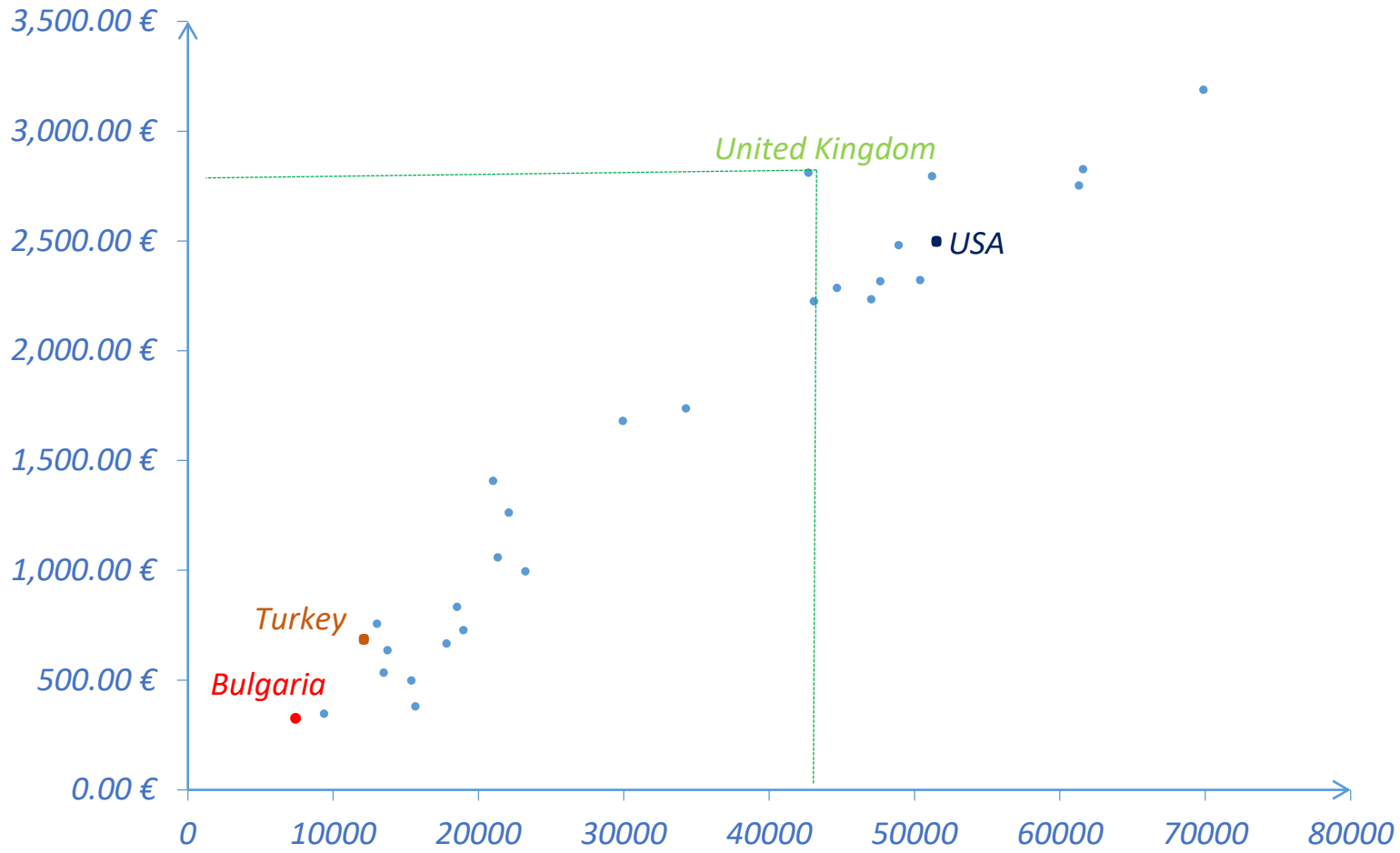
Фиг. 1 БНД на глава от населението (X) и разполагаем среден месечен доход на глава от населението (Y), през 2014 г. в Европейския съюз[1]



Фиг. 2 БНД на глава от населението (X) и разполагаем среден месечен доход на глава от населението (Y), през 2014 г. в Европейския съюз[1]



Фиг. 3 БНД на глава от населението (X) и разполагаем среден месечен доход на глава от населението (Y), през 2014 г. в Европейския съюз[1] +Турция



Фиг. 4 БНД на глава от населението (X) и разполагаем среден месечен доход на глава от населението (Y), през 2014 г. в Европейския съюз[1] +Турция и САЩ

Как бихте предложили да моделираме тази зависимост?

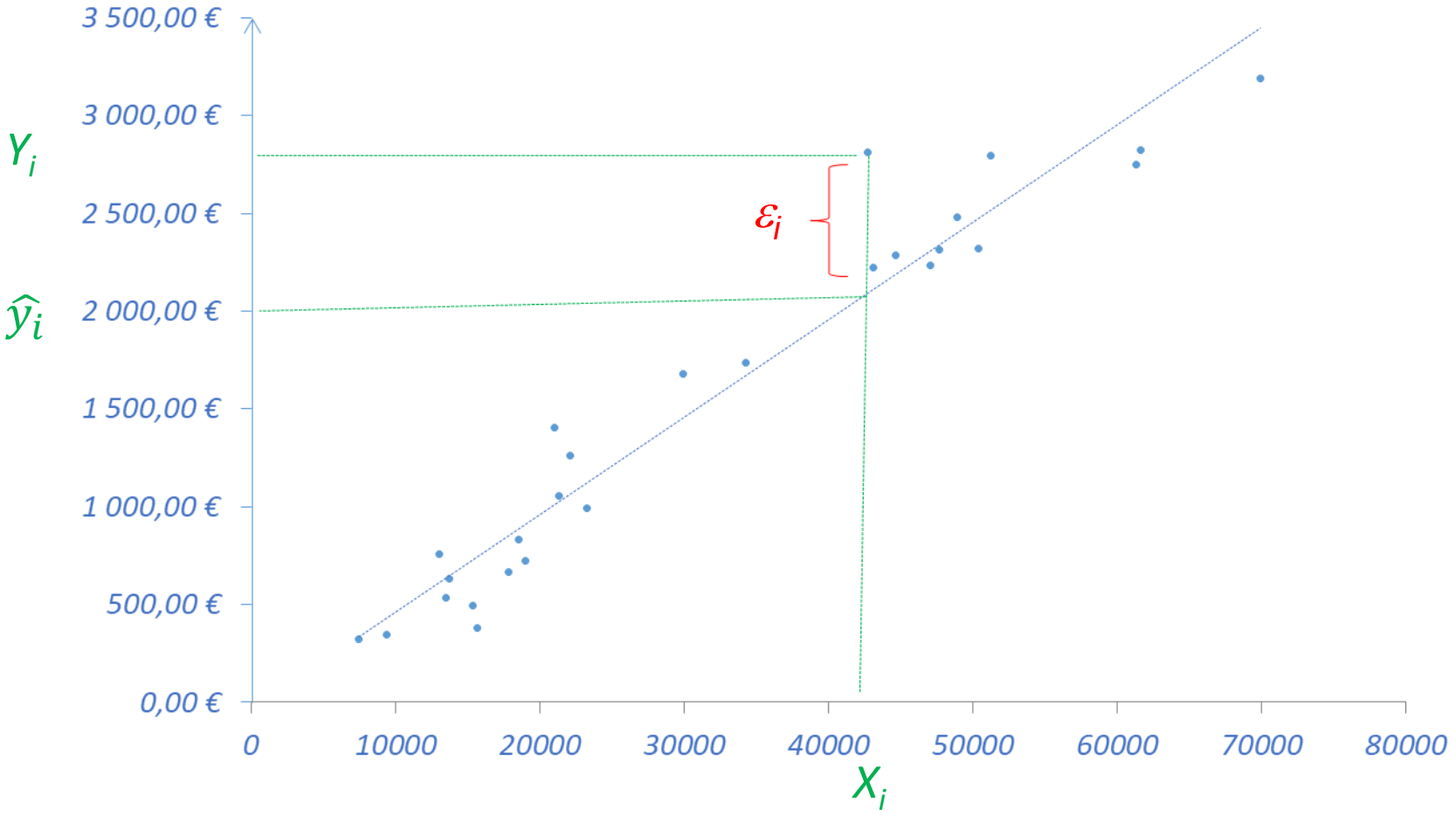
Как бихме могли да кажем при даден БНД какъв е очакваният разполагаем среден месечен доход на глава от населението?

Виждаме, че точките се групират около права.

Представяме я аналитично

(4) $$y = a_1 + a_2 x,$$

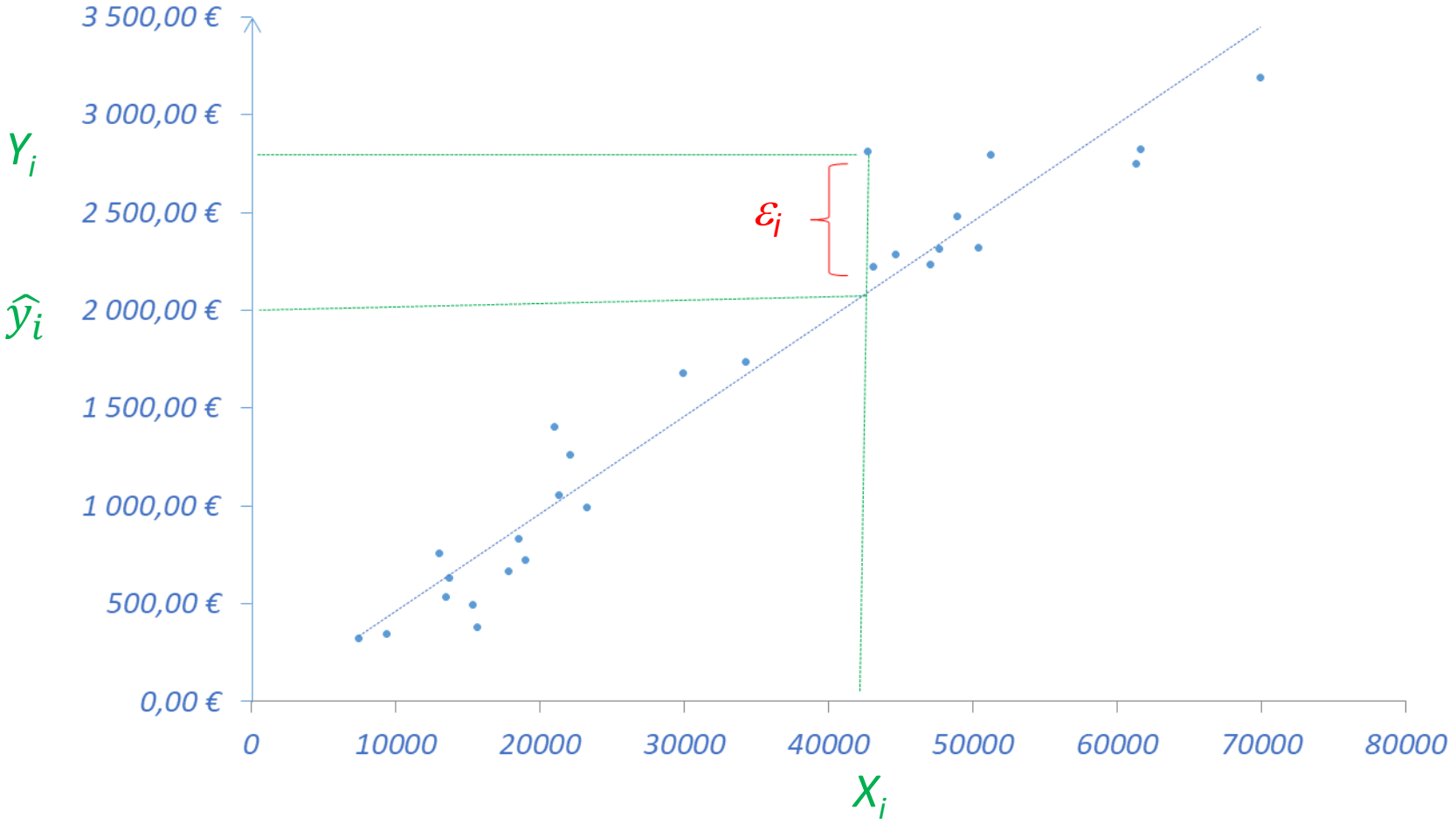
където a_1 *и* a_2 *са неизвестни параметри.*



Регресионният модел е

$$Y = a_1 + a_2 X + \varepsilon,$$

където a_1 *и* a_2 *са неизвестни параметри, а* ε *е стохастичната грешка на модела, която трябва да е независима от X, нормално разпределена и такава, че* $E\,\varepsilon = 0$, $D\,\varepsilon = \sigma^2$.



Неизвестните коефициенти a_1 *и* a_2 *в*

$$\hat{y} = a_1 + a_2 x,$$

се определят така, че да минимизират сумата от квадратите на грешките, т.е.

$$\underset{a_1, a_2}{argmin} \sum_{i=1}^{n} \varepsilon_i^2$$

$$\underset{a_1, a_2}{argmin} \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2$$

$$\underset{a_1, a_2}{argmin} \sum_{i=1}^{n} (Y_i - a_1 - a_2 x)^2$$

За да ги определим по данните от извадката, използваме метода на най-малките квадрати.

Т.е. построяваме оценка на вектора а = (a_1 , a_2) като решение на системата нормални уравнения

$$\begin{cases} \sum_{i=1}^{n} Y_i = a_1 n + a_2 \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} Y_i X_i = a_1 \sum_{i=1}^{n} X_i + a_2 \sum_{i=1}^{n} X_i^2 \end{cases}$$

Да означим това решение с $(\hat{a}_1, \hat{a}_2)$.

Тогава оценката на уравнението на линията на регресия е

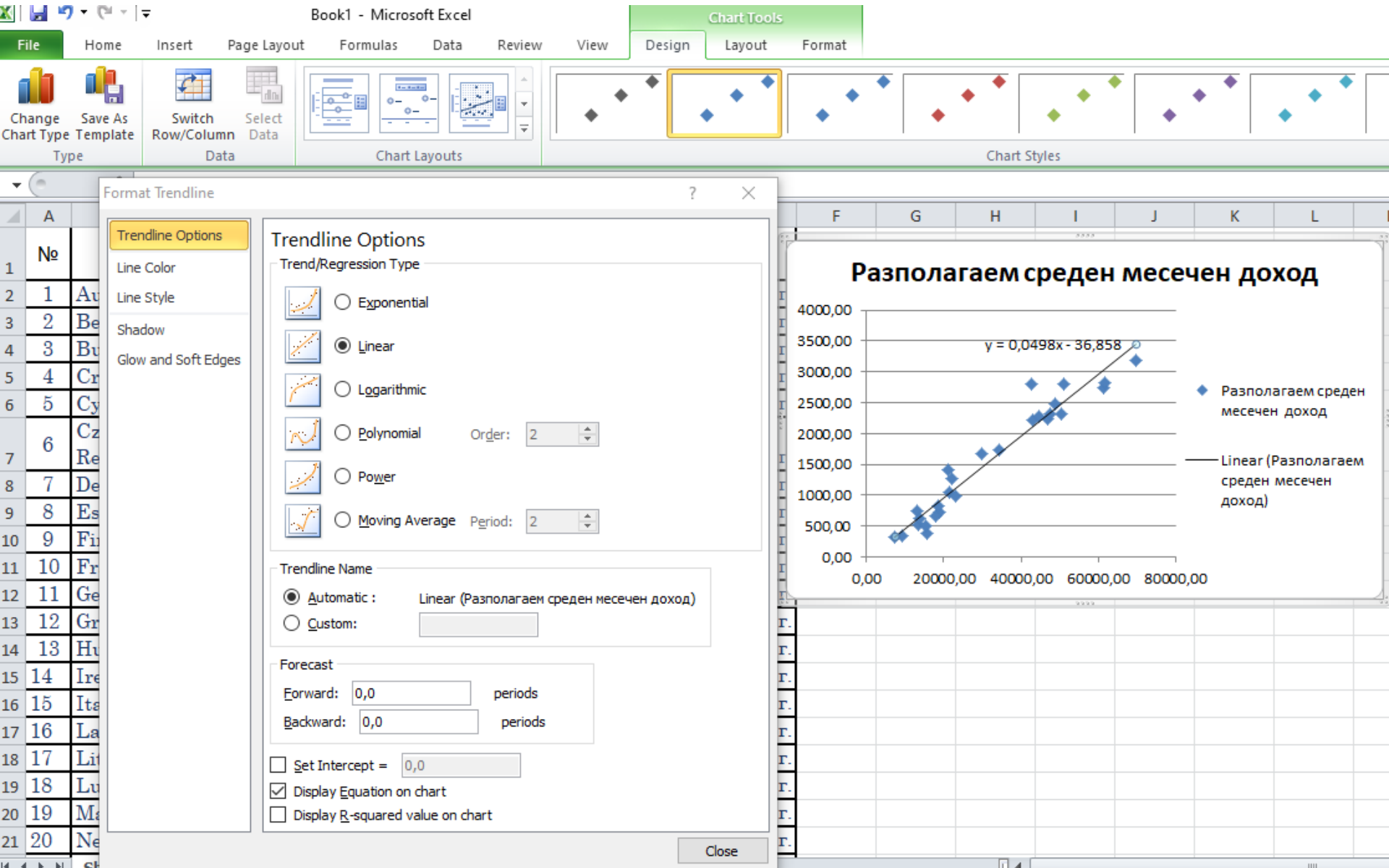
$$\hat{y} = \hat{a}_1 + \hat{a}_2 x.$$

И съответно регресионният модел е

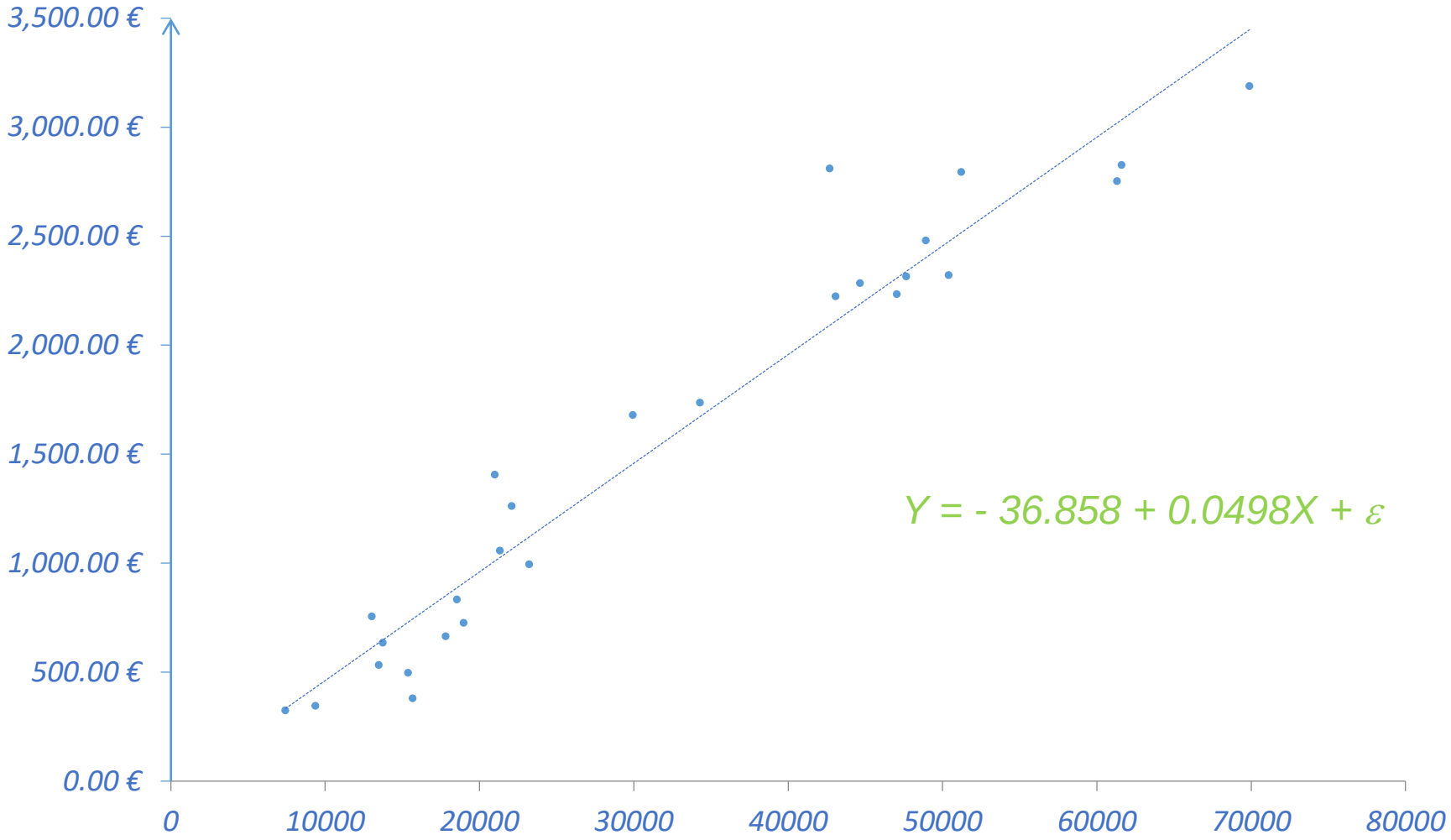
$$\hat{y} = \hat{a}_1 + \hat{a}_2 x + \varepsilon.$$

С помощта на Excel тази система се решава например като натиснем десен бутон на мишката върху точка от корелационното поле и от падащото меню изберем

Add Trendline (Добави линия на тренда), след което избираме линия на тренда и слагаме ъгълче на Display Equation on chart (Покажи уравнението на графиката).



След форматиране по избор получаваме например



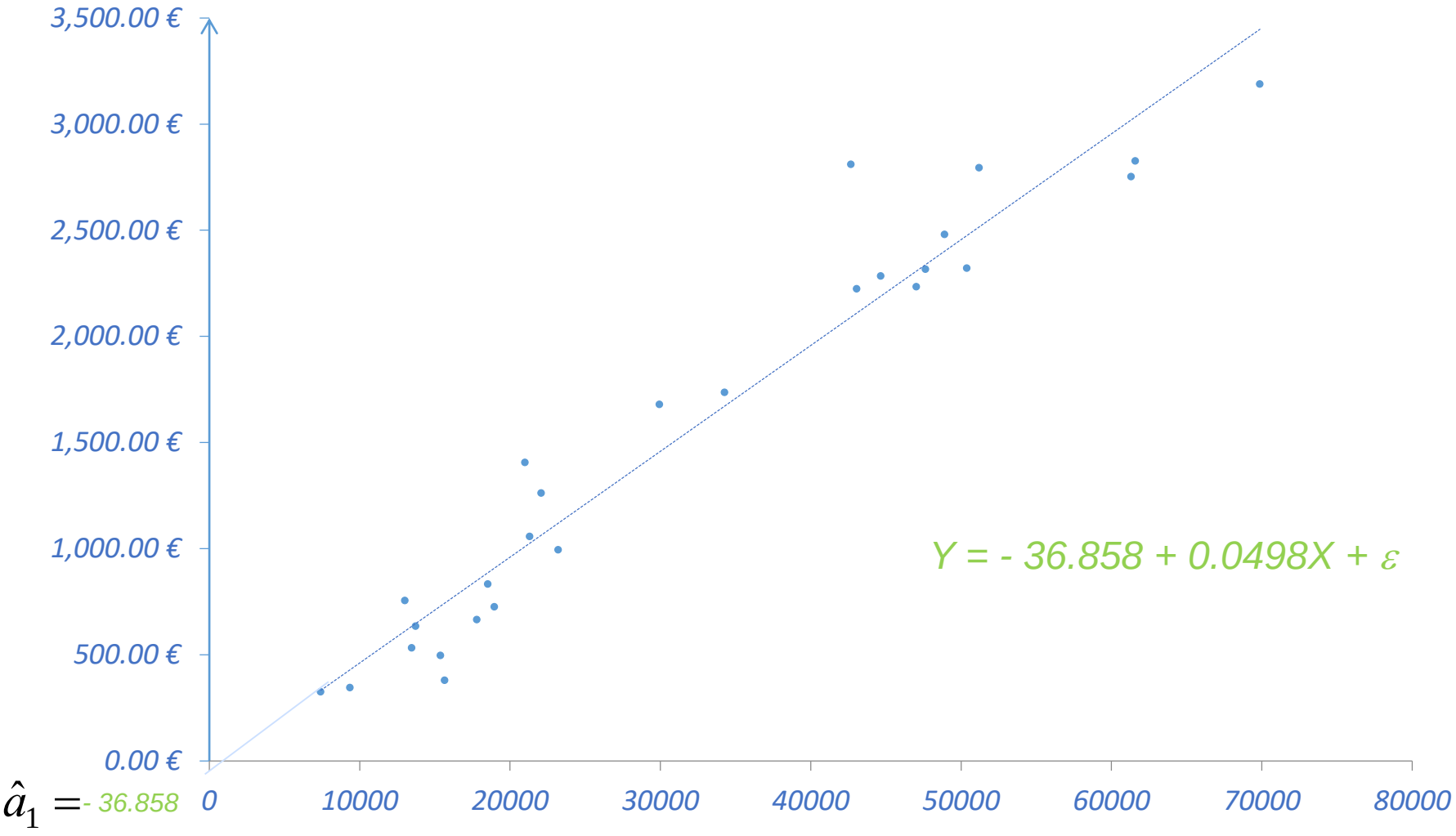
Коефициентът $\hat{a}_2$ *в това уравнение показва, с колко единици, в приетата за резултативния признак Y мярка, средно би се изменил той, ако изменим факторпризнака X с една единица в приетата за него мярка.*

В случая получихме, че когато БНД на глава от населението се увеличи с 1$, се очаква разполагаемият среден месечен доход на лице от населението в страните от генералната съвкупност да се увеличи средно с 0,0489 Евра.

Когато зависимостта на резултативния признак от факторпризнака е правопропорционална, коефициентът $\hat{a}_2$ *е положителен. Обратно, ако тази зависимост е обратнопропорционална, този коефициент е отрицателен.*

Линията на регресия ще е успоредна на абсцисната ос ако значенията на резултативния признак не се влияят от тези на факторпризнака. Тогава коефициентът $\hat{a}_2$ *е 0.*

Коефициентът $\hat{a}_1$ *е равен на ординатата на точката, в която линията на регресия пресича ординатната ос.*



$\hat{a}_1 =$ - 36.858

Като заместим измерените значения на факторпризнака X, в уравнението на регресия, намираме съответните оценки $\hat{y}$ *за значенията на резултативния признак Y.*

Сумата и съответно средната аритметична на тези оценки е равна на съответната характеристика на изходните данни.

За да можем да съпоставим този модел с останалите регресионни модели с адитивна грешка пресмятаме общата стандартна грешка на модела по формула (3).

Често пъти резултатите от изследването се оформят в таблица от вида:

Източник на дисперсията	Сума от квадратите	Степени на свобода	Дисперсия	F - критерий
Регресия	$SS_D = \sum_{i=1}^{n} (\hat{y}_i - \overline{Y}_n)^2$	1	$S_D^2 = \frac{SS_D}{1}$	F емп = $\frac{S_D^2}{S_\varepsilon^2}$
Отклонение от регресията	$SS_\varepsilon = \sum_{i=1}^{n} (\hat{y}_i - Y_i)^2$	n – 2	$S_\varepsilon^2 = \frac{SS_\varepsilon}{n-2}$	
Общо:	$SS_Y = \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2$	n – 1	$S_Y^2 = \frac{SS_Y}{n-1}$	

Ако тази грешка е 0, значи имаме пълно съвпадение на изходните данни за Y с техните оценки $\hat{y}$. *Тогава между X и Y има функционална, и по-точно праволинейна зависимост.*

Тогава имаме 2 възможности

1 случай. Оценките на Y не се влияят от стойностите на X и точките от корелационното поле са върху права, успоредна на абсцисната ос. Тогава всички оценки на резултативния признак ще са равни помежду си и по тази причина ще са равни на своята средна аритметична и на средната аритметична на изходните данни за този признак. Така $S_Y^2 = 0$ *и регресионния анализ в този случай не е подходящ.*

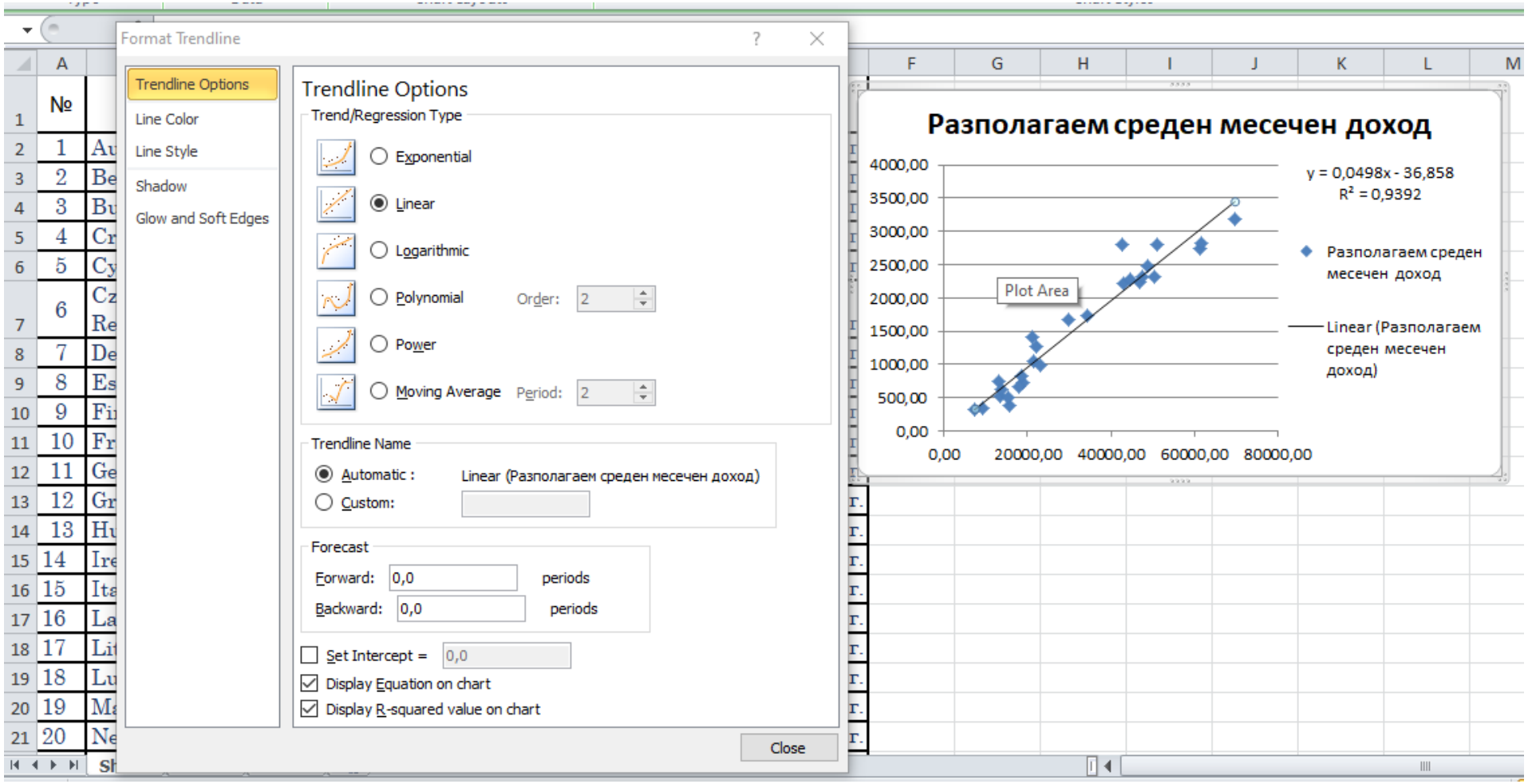
2 случай. Точките от корелационното поле са върху друга права. Тогава между наблюдаваните признаци има силна праволинейна зависимост и регресионния модел е много добър.

На основата на тези разсъждения е образуван корелационния коефициент на Пирсън, който в случая на праволинейна зависимост съвпада с корелационния коефициент на Браве

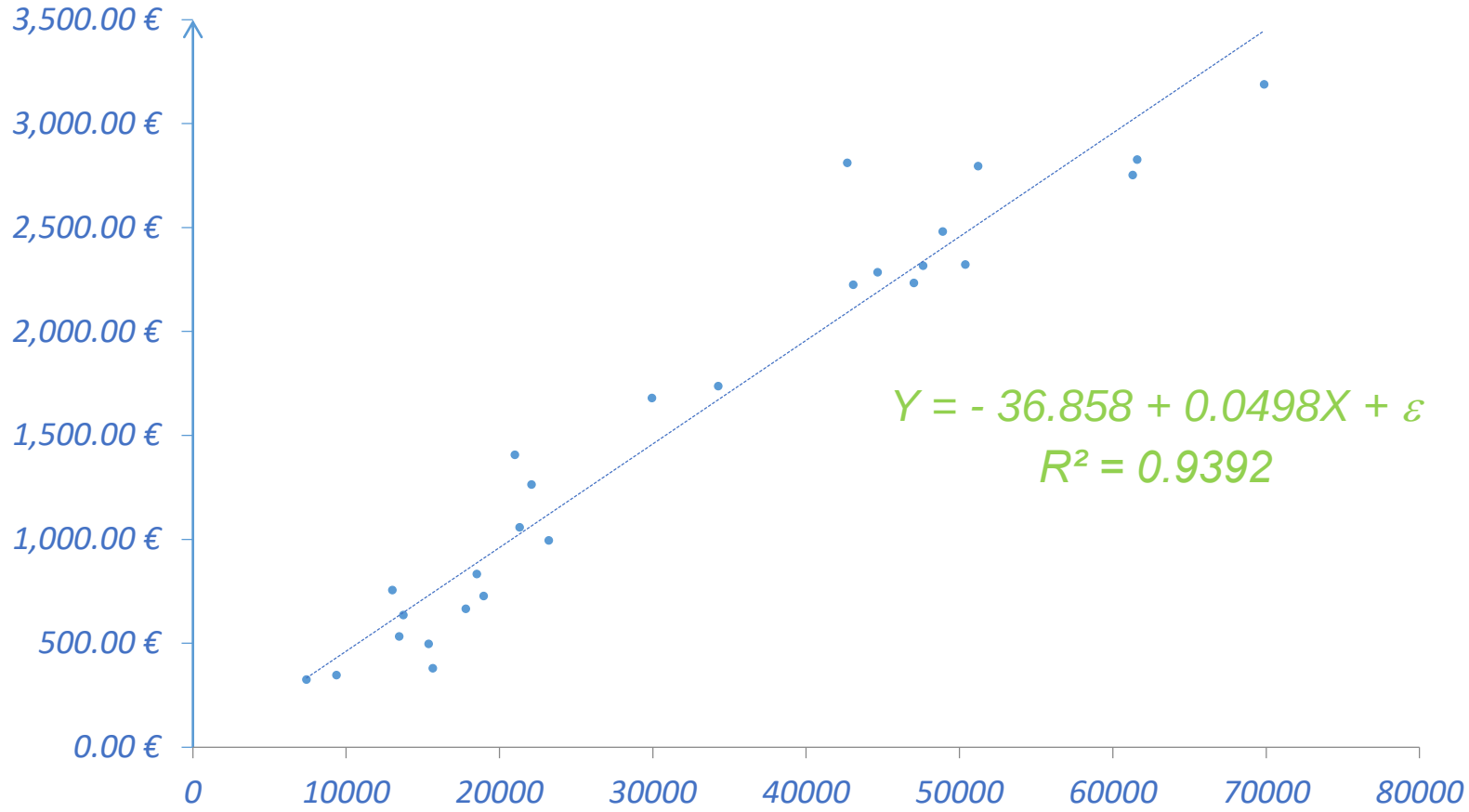
$$r = \pm\sqrt{1 - \frac{S_\varepsilon^2}{S_Y^2}}$$

Изменя се между -1 и 1 и се анализира както корелационния коефициент на Браве. В праволинейния модел знакът пред корена се определя от знака на коефициента пред независимата променлива. Когато е близо до 1 или -1 между наблюдаваните признаци съществува силна праволинейна зависимост. При r = +1 е правопропорционална и функционална. При -1 е обратнопропорционална и функционална.

С помощта на Excel квадрата му се получава например като натиснем десен бутон на мишката върху точка от корелационното поле и от падащото меню изберем Add Trendline (Добави линия на тренда), след което слагаме ъгълче на Display R-squared value on chart .



След форматиране по избор получаваме например



Както вече знаем r^2 е коефициента на детерминация. В случая 93,92% от вариацията на Y се дължат на изменения в X, т.е. X има силно влияние върху У. Статистиката, обаче не ни отговаря на въпроса дали това влияние е причинно следствено.

Проверяваме дали грешките са независими и нормално разпределени, дали са с нулево средно и с постоянна и крайна дисперсия.

Ако грешките са нормално разпределени следва проверка на хипотезата за адекватност на тествания модел, т.е. Дали наистина X оказва статистически значимо влияние на Y.

В случая проверяваме хипотезата

H_0: $a_2 = 0$, т.е. факторпризнакът X не оказва статистически значимо влияние на резултативната величина Y.

Алтернативата е

H_1: $a_2 \neq 0$.

Избираме риска за грешка от първи род $\alpha \in (0, 1)$.

Критичната област за нулевата хипотеза има вида

$$W_\alpha = \left\{ (x_1, \ldots, x_n) \in \Omega : \ \frac{S_D^2}{S_\varepsilon^2} \geq x_{1-\alpha, F(1, n-2)} \right\},$$

където $x_{1-\alpha, F(1,n-2)}$ *е 1-* α *квантилът на F(1, n - 2) разпределението.*

Вече можем да проверим допълнителни хипотези.

Например можем ли да закръглим кофициентите си. В този случай в числителя на емпиричната характеристика имаме разликата на оценката и тестватната константа, а в знаменателя стои стандартната грешка на оценката, която може да бъде намерена в следващата таблица. Костантата $x_{1-\alpha/2,t(n-2)}$ *е 1 - α /2 квантилът на t(n - 2).*

Можем да построим и доверителни интервали на a_1 *и* a_2 *и Y.*

Величина	Стандартна грешка	Степени на свобода	Граници на доверителния интервал
a_1	$S_{a_1} = S_\varepsilon \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}$	n – 2	$\hat{a}_1 \pm x_{1-\alpha/2,t(n-2)} S_{a_1}$
a_2	$S_{a_2} = S_\varepsilon \sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}$	n – 2	$\hat{a}_2 \pm x_{1-\alpha/2,t(n-2)} S_{a_2}$
Ако независимите променливи не са случайни	$S_{EY_i} = S_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X}_n)^2}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}$	n – 2	$\hat{y}_i \pm x_{1-\alpha/2,t(n-2)} S_{EY_i}$
Ако независимите променливи са случайни	$S_{EY/X_i=x_i} = S_\varepsilon \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X}_n)^2}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}$	n – 2	$\hat{y}_i \pm x_{1-\alpha/2,t(n-2)} S_{EY/X_i=x_i}$