Ποσοτικές Μέθοδοι στην Οικονομία και στη Διοίκηση 2

Ομαδική εργασία εξαμήνου

8220117 Ιωάννης Παπαδόπουλος

8220133 Μαρία – Ειρήνη Σαρμπάνη

8220210 Χαρίλαος Ντουράκης

8220154 Αικατερίνη Τσικούρα

8220035 Σπήλιος Δημακόπουλος

Θέμα 1.

Στο Worksheet ComponentLifetimes παρουσιάζεται η διάρκεια καλής λειτουργίας για 190 εξαρτήματα.

1. Υπολογίστε τη μέση τιμή και τη διάμεσο της διάρκειας καλής λειτουργίας.

2. Τι συμπεραίνουμε για τη συμμετρία της κατανομής;

3. Κατασκευάστε την Κατανομή Συχνοτήτων των Βαθμών. Ο αριθμός και το εύρος των διαστημάτων αποτελεί δική σας επιλογή.

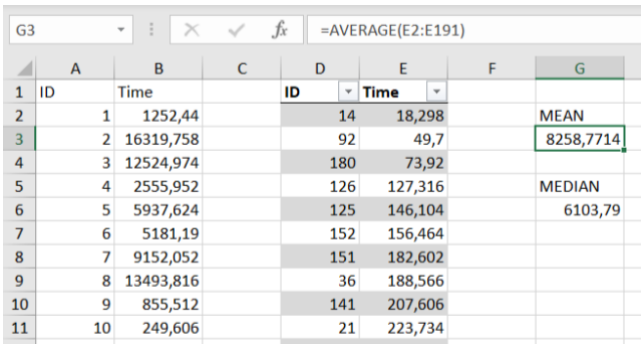
4. Σχεδιάστε το Ιστόγραμμα από τα δεδομένα της Κατανομής Συχνοτήτων.

5. Εκτιμήστε τη μέση τιμή και τη διάμεσο από τα δεδομένα της κατανομής Συχνοτήτων.

1. Υπολογίστε τη μέση τιμή και τη διάμεσο της διάρκειας καλής λειτουργίας.

Αν X είναι το χαρακτηριστικό το οποίο μετρά τη συνολική διάρκεια καλής λειτουργίας των εξαρτημάτων για 190 εξαρτήματα $X_1, X_2, X_3,\ldots, X_{190}$, τότε η ***Μέση Τιμή (MEAN)*** ορίζεται ως:

$$\mu_X = \frac{X_1+X_2+X_3+\cdots+X_{190}}{190} = \frac{1}{190} * \sum_{i=1}^{190} X_i = 8258{,}771389$$



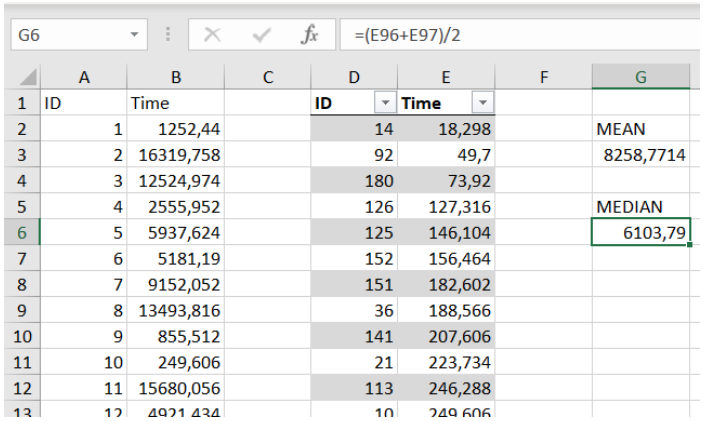
	A	B	C	D	E	F	G
1	ID	Time		ID	Time		
2	1	1252,44		14	18,298		MEAN
3	2	16319,758		92	49,7		8258,7714
4	3	12524,974		180	73,92		
5	4	2555,952		126	127,316		MEDIAN
6	5	5937,624		125	146,104		6103,79
7	6	5181,19		152	156,464		
8	7	9152,052		151	182,602		
9	8	13493,816		36	188,566		
10	9	855,512		141	207,606		
11	10	249,606		21	223,734		

Διάμεσος (MEDIAN)

Ο πίνακας στις στήλες D και E είναι τα δεδομένα μας ταξινομημένα με βάση τη διάρκεια καλής λειτουργίας των εξαρτημάτων

$\text{Θέση} : \text{M} = \frac{190+1}{2} = 95{,}5$ (ωστόσο στην περίπτωση αυτή δε βρίσκεται ο μέσος όρων μεταξύ των γραμμών 95 και 96 των στηλών E και D, αλλά 96 και 97 καθώς η 1η γραμμή είναι επικεφαλίδα στο στο excel μας)

Άρα $\frac{6098{,}848+6108{,}732}{2} = 6103{,}79$



G6 =(E96+E97)/2

	A	B	C	D	E	F	G
1	ID	Time		ID	Time		
2	1	1252,44		14	18,298		MEAN
3	2	16319,758		92	49,7		8258,7714
4	3	12524,974		180	73,92		
5	4	2555,952		126	127,316		MEDIAN
6	5	5937,624		125	146,104		6103,79
7	6	5181,19		152	156,464		
8	7	9152,052		151	182,602		
9	8	13493,816		36	188,566		
10	9	855,512		141	207,606		
11	10	249,606		21	223,734		
12	11	15680,056		113	246,288		
13	12	4921,434		10	249,606		

2. Τι συμπεραίνουμε για τη συμμετρία της κατανομής;

Εφόσον η Μέση Τιμή είναι διαφορετική της διαμέσου, η κατανομή δεν είναι κανονική και έχουμε ασυμμετρία.

Παρατηρούμε ότι Μέση Τιμή > Διάμεσος ($\mu_X = 8258{,}771389$) > (M =6103.79), άρα υπάρχει θετική ασυμμετρία, δηλαδή έχουμε ύπαρξη υψηλών ακραίων τιμών που «τραβούν» τη μέση τιμή δεξιά της διαμέσου.

3. Κατασκευάστε την Κατανομή Συχνοτήτων των Βαθμών. Ο αριθμός και το εύρος των διαστημάτων αποτελεί δική σας επιλογή.

Σύμφωνα με τον ταξινομημένο μας πίνακα, κατασκευάζουμε τον πίνακα κατανομής συχνοτήτων με 18 διαστήματα εύρους 2500 βαθμών ως εξής:

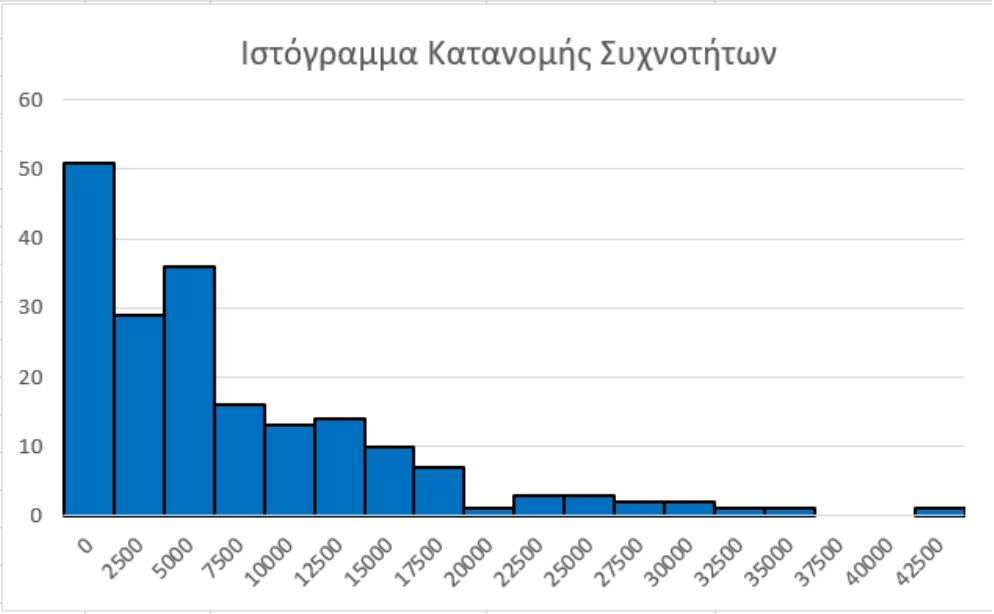
	Διαστήματα Τάξης		Συχνότητα	Σχετική Συχνότητα	Σχετική Αθροιστική Συχνότητα		
	Time		Αριθμός Εξαρτημάτων	Ποσοστό(%)	Αθροιστική Συχνότητα	Αθροιστική Συχνότητα(%)	Κεντρικός Όρος
0	και κάτω των	2500	51	26,84210526	51	26,84210526	1250
2500	και κάτω των	5000	29	15,26315789	80	42,10526316	3750
5000	και κάτω των	7500	36	18,94736842	116	61,05263158	6250
7500	και κάτω των	10000	16	8,421052632	132	69,47368421	8750
10000	και κάτω των	12500	13	6,842105263	145	76,31578947	11250
12500	και κάτω των	15000	14	7,368421053	159	83,68421053	13750
15000	και κάτω των	17500	10	5,263157895	169	88,94736842	16250
17500	και κάτω των	20000	7	3,684210526	176	92,63157895	18750
20000	και κάτω των	22500	1	0,526315789	177	93,15789474	21250
22500	και κάτω των	25000	3	1,578947368	180	94,73684211	23750
25000	και κάτω των	27500	3	1,578947368	183	96,31578947	26250
27500	και κάτω των	30000	2	1,052631579	185	97,36842105	28750
30000	και κάτω των	32500	2	1,052631579	187	98,42105263	31250
32500	και κάτω των	35000	1	0,526315789	188	98,94736842	33750
35000	και κάτω των	37500	1	0,526315789	189	99,47368421	36250
37500	και κάτω των	40000	0	0	189	99,47368421	38750
40000	και κάτω των	42500	0	0	189	99,47368421	41250
42500	και κάτω των	45000	1	0,526315789	190	100	
	ΣΥΝΟΛΟ		190				

4. Σχεδιάστε το Ιστόγραμμα από τα δεδομένα της Κατανομής Συχνοτήτων.

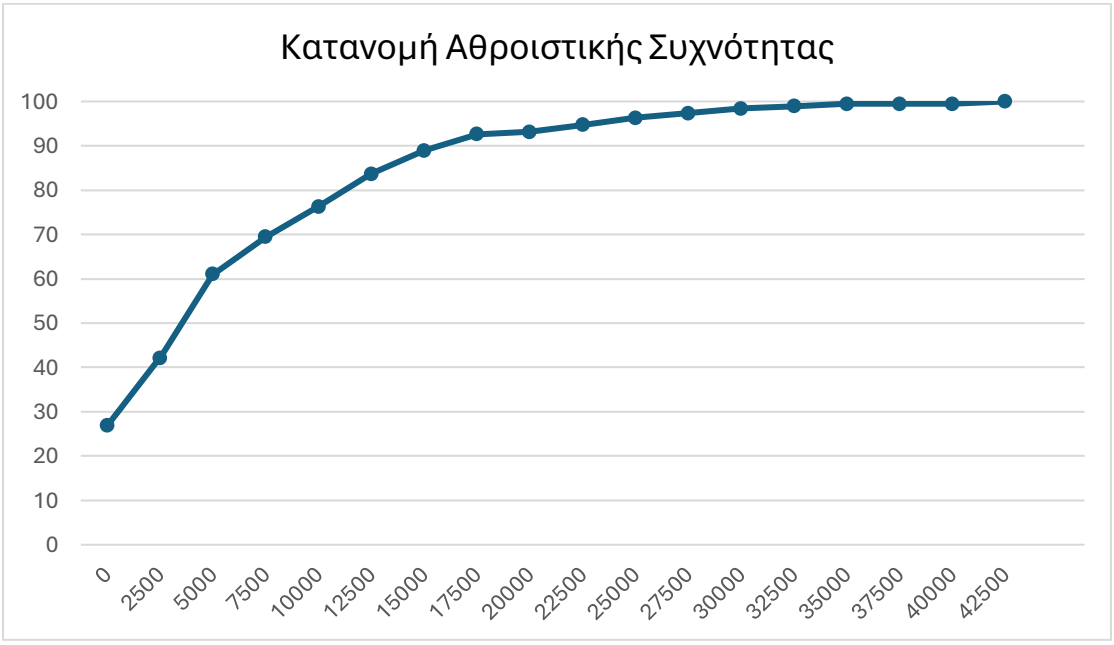
Ιστόγραμμα Κατανομής Συχνοτήτων

Υπολογίζουμε τους μέσους κάθε διαστήματος και τους τοποθετούμε στη βάση του διαγράμματος

Κεντρικός όρος
1250
3750
6250
8750
11250
13750
16250
18750
21250
23750
26250
28750
31250
33750
36250
38750
41250
43750



5. Εκτιμήστε τη μέση τιμή και τη διάμεσο από τα δεδομένα της κατανομής Συχνοτήτων.



	Διαστήματα Τάξης		Συχνότητα	Σχετική Συχνότητα	Σχετική Αθροιστική Συχνότητα		
	Time		Αριθμός Εξαρτημάτων	Ποσοστό(%)	Αθροιστική Συχνότητα	Αθροιστική Συχνότητα(%)	Κεντρικός Όρος
0	και κάτω των	2500	51	26,84210526	51	26,84210526	1250
2500	και κάτω των	5000	29	15,26315789	80	42,10526316	3750
5000	και κάτω των	7500	36	18,94736842	116	61,05263158	6250
7500	και κάτω των	10000	16	8,421052632	132	69,47368421	8750
10000	και κάτω των	12500	13	6,842105263	145	76,31578947	11250
12500	και κάτω των	15000	14	7,368421053	159	83,68421053	13750
15000	και κάτω των	17500	10	5,263157895	169	88,94736842	16250
17500	και κάτω των	20000	7	3,684210526	176	92,63157895	18750
20000	και κάτω των	22500	1	0,526315789	177	93,15789474	21250
22500	και κάτω των	25000	3	1,578947368	180	94,73684211	23750
25000	και κάτω των	27500	3	1,578947368	183	96,31578947	26250
27500	και κάτω των	30000	2	1,052631579	185	97,36842105	28750
30000	και κάτω των	32500	2	1,052631579	187	98,42105263	31250
32500	και κάτω των	35000	1	0,526315789	188	98,94736842	33750
35000	και κάτω των	37500	1	0,526315789	189	99,47368421	36250
37500	και κάτω των	40000	0	0	189	99,47368421	38750
40000	και κάτω των	42500	0	0	189	99,47368421	41250
42500	και κάτω των	45000	1	0,526315789	190	100	
	ΣΥΝΟΛΟ		190				

Εκτίμηση Μέσης τιμής:

$\mu_X = \frac{\sum_{i=1}^{18} xi*vi}{\sum_{i=1}^{18} vi}$ όπου xi ο κεντρικός όρος και vi η συχνότητα

Άρα $\mu_X = \frac{1531250}{190} = 8059{,}210526$

Εκτίμηση Διαμέσου:

Η Διάμεσος σύμφωνα με τη γραμμική παρεμβολή υπολογίζεται ως εξής:

$$M = L_M + \delta * \frac{(\frac{n}{2} - F_{M-1})}{f_M}$$

Όπου

L_M : Το τελευταίο σημείο με αθροιστική κατανομή κάτω από 50%, αλλιώς το κάτω όριο του διαστήματος (Μ) που εντοπίζεται η διάμεσος, δηλαδή 5000

δ: Το πλάτος ενός διαστήματος, δηλαδή 2500

n: Το μέγεθος του πληθυσμού, δηλαδή 190

F_{M-1}: Η αθροιστική κατανομή του διαστήματος πριν το διάστημα Μ, δηλαδή 80

f_M: Η συχνότητα του διαστήματος Μ που περιέχει τη διάμεσο, δηλαδή 36

Αντικαθιστώντας έχουμε:

$$M = 5000 + 2500 * \frac{\left(\frac{190}{2} - 80\right)}{36} = 5000 + 2500 * \frac{15}{36} \approx 6041{,}66$$

Θέμα 2.

Στο Worksheet PopulationHeight παρουσιάζεται το ύψος ενός πληθυσμού έξι ατόμων.

1. Υπολογίστε τη μέση τιμή, τη διακύμανση και την τυπική απόκλιση του ύψους του πληθυσμού.

2. Προσδιορίστε την κατανομή δειγματοληψίας με επανατοποθέτηση του ύψους για μέγεθος δείγματος 2.

3. Ποια είναι η μέση τιμή του δειγματικού μέσου;

4. Ποια είναι η διακύμανση του δειγματικού μέσου;

5. Ποια είναι η μέση τιμή των δειγματικών διακυμάνσεων; Ταυτίζεται με τη διακύμανση του πληθυσμού;

6. Ποια είναι η μέση τιμή των δειγματικών τυπικών αποκλίσεων; Ταυτίζεται με την τυπική απόκλιση του πληθυσμού;

7. Αν λάβουμε δείγμα μεγέθους 30 (με επανατοποθέτηση), ποια η πιθανότητα ο δειγματικός μέσος να είναι πάνω από 179 cm;

1. Υπολογίστε τη μέση τιμή, τη διακύμανση και την τυπική απόκλιση του ύψους του πληθυσμού.

Ατομο	**Height**
1	182
2	174
3	198
4	175
5	161
6	176

Μέση τιμή:

$$\mu = \frac{1}{v}\sum_{i=1}^{v} x_i = \frac{1}{6}\sum_{i=1}^{6}(182 + 174 + 198 + 175 + 161 + 176) = 177{,}66666667$$

Διακύμανση:

$$\sigma^2 = \frac{1}{v}\sum_{i=1}^{v}(x_i - \mu)^2 = \frac{1}{6}\sum_{i=1}^{6}[(182 - 177{,}66666667)^2 + (174 - 177{,}66666667)^2 + (198 - 177{,}66666667)^2 + (175 - 177{,}66666667)^2 + (161 - 177{,}66666667)^2 + (176 - 177{,}66666667)^2] = 122{,}222222$$

Τυπική απόκλιση:

$$\sigma = \sqrt{\sigma^2} = = \sqrt{122{,}2222} = 11{,}5541597$$

2. Προσδιορίστε την κατανομή δειγματοληψίας με επανατοποθέτηση του ύψους για μέγεθος δείγματος 2.

Κατανομή δειγματοληψίας με επανατοποθέτηση για πληθυσμό v=6 και μέγεθος δείγματος 2, άρα θα έχουμε $6^2 = 36$ δείγματα

			τιμες παρατηρήσεων		
Δειγμα	Επιλογή 1	Επιλογή 2	X1	X2	ΔΕΙΓΜ ΜΕΣΟΣ
1	1	1	182	182	182
2	1	2	182	174	178
3	1	3	182	198	190
4	1	4	182	175	178,5
5	1	5	182	161	171,5
6	1	6	182	176	179
7	2	1	174	182	178
8	2	2	174	174	174
9	2	3	174	198	186
10	2	4	174	175	174,5
11	2	5	174	161	167,5
12	2	6	174	176	175
13	3	1	198	182	190
14	3	2	198	174	186
15	3	3	198	198	198
16	3	4	198	175	186,5
17	3	5	198	161	179,5
18	3	6	198	176	187
19	4	1	175	182	178,5
20	4	2	175	174	174,5
21	4	3	175	198	186,5
22	4	4	175	175	175
23	4	5	175	161	168
24	4	6	175	176	175,5
25	5	1	161	182	171,5
26	5	2	161	174	167,5
27	5	3	161	198	179,5
28	5	4	161	175	168
29	5	5	161	161	161
30	5	6	161	176	168,5
31	6	1	176	182	179
32	6	2	176	174	175
33	6	3	176	198	187
34	6	4	176	175	175,5
35	6	5	176	161	168,5
36	6	6	176	176	176

3. Ποια είναι η μέση τιμή του δειγματικού μέσου;

Η μέση τιμή του δειγματικού μέσου είναι ίση με τη μέση τιμή του πληθυσμού και έχουμε :

$$\mu_{\bar{x}} = E(\bar{X}) = \mu = 177{,}66666667$$

Όπου $\mu_{\bar{x}}$ η μέση τιμή του δειγματικού μέσου

Και μ η μέση τιμή του πληθυσμού

4. Ποια είναι η διακύμανση του δειγματικού μέσου;

Διακύμανση του πληθυσμού των δειγματικών μέσων $\sigma_{\bar{x}}^2$ (Διακύμανσή Κατανομής Δειγματοληψίας)

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{122{,}2222222}{2} = 61{,}111111$$

5. Ποια είναι η μέση τιμή των δειγματικών διακυμάνσεων; Ταυτίζεται με τη διακύμανση του πληθυσμού;

			τιμες παρατηρήσεων			
Δειγμα	Επιλογή 1	Επιλογή 2	X1	X2	ΔΕΙΓΜ ΜΕΣΟΣ	δειγμ διακυμανσεις
1	1	1	182	182	182	0
2	1	2	182	174	178	32
3	1	3	182	198	190	128
4	1	4	182	175	178,5	24,5
5	1	5	182	161	171,5	220,5
6	1	6	182	176	179	18
7	2	1	174	182	178	32
8	2	2	174	174	174	0
9	2	3	174	198	186	288
10	2	4	174	175	174,5	0,5
11	2	5	174	161	167,5	84,5
12	2	6	174	176	175	2
13	3	1	198	182	190	128
14	3	2	198	174	186	288
15	3	3	198	198	198	0
16	3	4	198	175	186,5	264,5
17	3	5	198	161	179,5	684,5
18	3	6	198	176	187	242
19	4	1	175	182	178,5	24,5
20	4	2	175	174	174,5	0,5
21	4	3	175	198	186,5	264,5
22	4	4	175	175	175	0
23	4	5	175	161	168	98
24	4	6	175	176	175,5	0,5
25	5	1	161	182	171,5	220,5
26	5	2	161	174	167,5	84,5
27	5	3	161	198	179,5	684,5
28	5	4	161	175	168	98
29	5	5	161	161	161	0
30	5	6	161	176	168,5	112,5
31	6	1	176	182	179	18
32	6	2	176	174	175	2
33	6	3	176	198	187	242
34	6	4	176	175	175,5	0,5
35	6	5	176	161	168,5	112,5
36	6	6	176	176	176	0
					Μέση Τιμή Δειγμ. Διακυμάνσεων =	122,2222222

Η μέση τιμή των δειγματικών διακυμάνσεων είναι:

$$\mathrm{E}\left(\sigma_{\bar{x}}^2\right) = \frac{1}{36}\sum_{1}^{36}\sigma_{\bar{x}_i}^2 = 122{,}222222 = \sigma^2$$

Άρα η μέση τιμή των δειγματικών διακυμάνσεων είναι ίδια με αυτή του πληθυσμού.

6. Ποια είναι η μέση τιμή των δειγματικών τυπικών αποκλίσεων; Ταυτίζεται με την τυπική απόκλιση του πληθυσμού;

			τιμες παρατηρήσεων			
Δειγμα	Επιλογή 1	Επιλογή 2	X1	X2	ΔΕΙΓΜ ΜΕΣΟΣ	τυπ αποκλισεις
1	1	1	182	182	182	0
2	1	2	182	174	178	5,656854249
3	1	3	182	198	190	11,3137085
4	1	4	182	175	178,5	4,949747468
5	1	5	182	161	171,5	14,8492424
6	1	6	182	176	179	4,242640687
7	2	1	174	182	178	5,656854249
8	2	2	174	174	174	0
9	2	3	174	198	186	16,97056275
10	2	4	174	175	174,5	0,707106781
11	2	5	174	161	167,5	9,192388155
12	2	6	174	176	175	1,414213562
13	3	1	198	182	190	11,3137085
14	3	2	198	174	186	16,97056275
15	3	3	198	198	198	0
16	3	4	198	175	186,5	16,26345597
17	3	5	198	161	179,5	26,1629509
18	3	6	198	176	187	15,55634919
19	4	1	175	182	178,5	4,949747468
20	4	2	175	174	174,5	0,707106781
21	4	3	175	198	186,5	16,26345597
22	4	4	175	175	175	0
23	4	5	175	161	168	9,899494937
24	4	6	175	176	175,5	0,707106781
25	5	1	161	182	171,5	14,8492424
26	5	2	161	174	167,5	9,192388155
27	5	3	161	198	179,5	26,1629509
28	5	4	161	175	168	9,899494937
29	5	5	161	161	161	0
30	5	6	161	176	168,5	10,60660172
31	6	1	176	182	179	4,242640687
32	6	2	176	174	175	1,414213562
33	6	3	176	198	187	15,55634919
34	6	4	176	175	175,5	0,707106781
35	6	5	176	161	168,5	10,60660172
36	6	6	176	176	176	0
					Μέση Τιμή Δειγματικών Τυπικών Αποκλίσεων =	8,249579114

Η μέση τιμή των δειγματικών τυπικών αποκλίσεων είναι :

$$\frac{1}{36}\sum_{1}^{36}\sigma_{\bar{x}_i} = 8{,}249579114 \neq \sigma$$

Άρα δεν ταυτίζεται με την τυπική απόκλιση του πληθυσμού.

7. Αν λάβουμε δείγμα μεγέθους 30 (με επανατοποθέτηση), ποια η πιθανότητα ο δειγματικός μέσος να είναι πάνω από 179 cm;

Επειδή το δείγμα n=30 είναι μεγάλο, ακολουθείται η κανονική κατανομή με:

μέση τιμή μ_x= 177,66666667

διακύμανση $\sigma_x^2 = \frac{\sigma^2}{n} = \frac{122{,}222222}{30} = 4{,}074074$

τυπική απόκλιση $\sigma_x = \sqrt{\sigma_x^2} = 2{,}018433568$

Η πιθανότητα ο δειγματικός μέσος να είναι πάνω από 179 είναι

$$\mathrm{P}(\mathrm{X} > 179) = 1 - P(X < 179) = 1 - P\left(\frac{X-\mu_x}{\sigma_x} < \frac{179 - 177{,}66666667}{2{,}018433568}\right) = 1 - P\left(\frac{X-\mu_x}{\sigma_x} < 0{,}66\right)$$
$$= 1 - \Phi(0{,}66) = 1 - 0{,}7454 = 0{,}2546$$

Μέσω συνάρτησης excel βγαίνει 0,254441407 ακριβώς.

Θέμα 3.

Ο Πρόεδρος του ΔΕΤ θέλει να διεξάγει μία έρευνα για τον προσδιορισμό του ποσοστού των αποφοίτων του τμήματος, οι οποίοι βρήκαν εργασία εντός τριών μηνών από την αποφοίτησή τους. Για το λόγο αυτό σας αναθέτει τη διεξαγωγή μίας έρευνας και μας ζητάει η εκτίμηση διαστήματος του ποσοστού των αποφοίτων οι οποίοι απασχολήθηκαν εντός τριών μηνών από τη λήψη του πτυχίου τους να είναι ±2% σε επίπεδο εμπιστοσύνης 98%.

1.Ποιο είναι το απαιτούμενο μέγεθος δείγματος για τη ζητούμενη ακρίβεια;

2. Αν η έρευνα πραγματοποιηθεί στο μέγεθος πληθυσμού που προσδιορίστηκε στο ερώτημα Α και έχει σαν αποτέλεσμα το 78% των ερωτηθέντων να έχει βρει εργασία εντός τριμήνου από την αποφοίτηση, ποιο είναι το διάστημα εμπιστοσύνης για το ποσοστό στον πληθυσμό των αποφοίτων; Συγκρίνετε το εύρος του διαστήματος με το εύρος που είχε ζητήσει ο Πρόεδρος του Τμήματος. Σχολιάστε τα αποτελέσματα της σύγκρισης.

1.Ποιο είναι το απαιτούμενο μέγεθος δείγματος για τη ζητούμενη ακρίβεια;

Εκτίμηση διαστήματος: **e = ± 2% = ± 0,02** επιτρεπόμενο σφάλμα (ακρίβεια)

Επίπεδο εμπιστοσύνης: 1 – α = 98% ⇔ 1 - α = 0,98 ⇔ **α = 0,02** περιθώριο σφάλματος

Για να υπολογίσουμε το απαιτούμενο μέγεθος δείγματος για τη ζητούμενη ακρίβεια χρησιμοποιούμε τον τύπο:

$$n = \frac{{Z_{a/2}}^2 \cdot p \cdot (1-p)}{e^2}$$

Όπου, n: το μέγεθος του δείγματος, p: το ποσοστό του πληθυσμού που έχει το χαρακτηριστικό που μας ενδιαφέρει (οι φοιτητές που απασχολήθηκαν εντός τριών μηνών από τη λήψη του πτυχίου τους), e: ακρίβεια και α: το περιθώριο σφάλματος.

Επειδή δε γνωρίζουμε το ποσοστό p, θέτουμε p = 0,5 καθώς είναι η τιμή η οποία μας μεγιστοποιεί το μέγεθος του χωρίς να ξεφεύγουμε από την δοσμένη ακρίβεια.

Άρα έχουμε,

$$n = \frac{{Z_{0,02/2}}^2 * 0,5 * (1-0,5)}{0,02^2} = \frac{{Z_{0,01}}^2 * 0,5 * 0,5}{0,02^2} = \frac{2,33^2 * 0,5^2}{0,02^2} = 3393$$

Επομένως, απαιτούμενο μέγεθος δείγματος για τη ζητούμενη ακρίβεια είναι **n =3393**.

2. Αν η έρευνα πραγματοποιηθεί στο μέγεθος πληθυσμού που προσδιορίστηκε στο ερώτημα Α και έχει σαν αποτέλεσμα το 78% των ερωτηθέντων να έχει βρει εργασία εντός τριμήνου από την αποφοίτηση, ποιο είναι το διάστημα εμπιστοσύνης για το ποσοστό στον πληθυσμό των αποφοίτων; Συγκρίνετε το εύρος του διαστήματος με το εύρος που είχε ζητήσει ο Πρόεδρος του Τμήματος. Σχολιάστε τα αποτελέσματα της σύγκρισης.

$\hat{p}$ = 78% = 0,78 το ποσοστό του δείγματος που έχει το χαρακτηριστικό που μας ενδιαφέρει (οι φοιτητές που απασχολήθηκαν εντός τριών μηνών από τη λήψη του πτυχίου τους)

η τυπική απόκλιση του ποσοστού δίνεται από τον τύπο: $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Όπου, $\sigma_{\hat{p}}$: η τυπική απόκλιση, $\hat{p}$: το ποσοστό του δείγματος που έχει το χαρακτηριστικό που μας ενδιαφέρει και n: το μέγεθος του δείγματος.

$$\sigma_{\hat{p}} = \sqrt{\frac{0{,}78(1 - 0{,}78)}{3393}} = \sqrt{\frac{0{,}78 * 0{,}22}{3393}} = 0{,}00711$$

Άρα, $\boldsymbol{\sigma_{\hat{p}} = 0{,}0711}$ και το διάστημα εμπιστοσύνης δίνεται από τον τύπο:

$[\ \hat{p} - Z_{0,02/2} * \sigma_{\hat{p}},\ \hat{p} + Z_{0,02/2} * \sigma_{\hat{p}}\ ]$

$[\ 0{,}78 - Z_{0,01} * 0{,}00711\ ,\ 0{,}78 + Z_{0,01} * 0{,}00711\ ] =$

$[\ 0{,}78 - 2{,}33 * 0{,}00711\ ,\ 0{,}78 + 2{,}33 * 0{,}00711\ ] = [\ 0{,}78 - 0{,}016\ ,\ 0{,}78 + 0{,}016\ ]$

Επομένως, αφού το εύρος διαστήματος (σφάλμα): $\pm$ 0,016 είναι μικρότερο από το δοσμένο e = $\pm$ 0,02, η έρευνα μας με μέγεθος δείγματος n = 3393 τηρεί τις προδιαγραφές που μας είχε ζητήσει ο Πρόεδρος του Τμήματος.

Θέμα 4.

Έχουμε στην κατοχή μας ποσοτικά δεδομένα ενός δείγματος 50 ατόμων από ένα πληθυσμό, ο οποίος έχει διακύμανση σ^2 = 121. Διεξάγεται ο παρακάτω έλεγχος: $H0$: $\mu \leq 100$ $H1$: $\mu > 100$

1. Υπολογίστε την πιθανότητα σφάλματος τύπου II, αν ο πραγματικός μέσος είναι μ = 100.5 σε επίπεδο σημαντικότητας α = 0.2

2. Υπολογίστε την πιθανότητα σφάλματος τύπου II, αν ο πραγματικός μέσος είναι μ = 100.5 σε επίπεδο σημαντικότητας α = 0.02

Συγκρίνετε τα αποτελέσματα των ερωτημάτων 1 και 2 και σχολιάστε τι προκύπτει από τη σύγκριση.

1. Υπολογίστε την πιθανότητα σφάλματος τύπου II, αν ο πραγματικός μέσος είναι μ = 100.5 σε επίπεδο σημαντικότητας α = 0.2

Διατύπωση Υποθέσεων

H_0: μ ≤ 100

H_1: μ > 100

Ο έλεγχος που πρέπει να διεξαχθεί είναι μονοκατάληκτος προς τα πάνω

Άρα η μεταβλητή $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ στο άνω όριο έχει οριακή τιμή :

Πάνω Όριο: $z_{(1-\alpha)} = z_{(1-0,2)} = Z_{0.8} = 0.84$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{121}}{\sqrt{50}} = \frac{11}{7.071} = 1.555649 \approx \mathbf{1,55}$$

Όταν έχουμε σφάλμα τύπου || σημαίνει ότι έχουμε αποδεχτεί την $H0$ χωρίς να ισχύει. Επομένως , η μεταβλητή Z θα πρέπει να βρίσκεται εντός της ζώνης αποδοχής.

Δηλαδή, πρέπει : $\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < 0,84 \Rightarrow \frac{\bar{X}-100}{1,55} * 1,55 < 0,84 * 1,55 \Rightarrow \bar{X} - 100 < 1,302 \Rightarrow \bar{X} - 100 + 100 < 1,302 + 100 \Rightarrow \bar{X} < 101,302$

Δίνεται ο μέσος του πληθυσμού μ=100.5

Σφάλμα Τύπου || Λανθασμένη μη απόρριψη, δηλαδή να μην απορρίψουμε λανθασμένα την H0

$$\beta = P\{\bar{X} < 101.302\} = P\left\{\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{101.302-100,5}{1.55}\right\} = P\left\{\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{0,802}{1.55}\right\} = \Phi(0,52) = 0,6985$$

Άρα, η πιθανότητα σφάλματος τύπου || σε επίπεδο σημαντικότητας α=0,2 είναι Β=0,6985.

2. Υπολογίστε την πιθανότητα σφάλματος τύπου II, αν ο πραγματικός μέσος είναι μ = 100.5 σε επίπεδο σημαντικότητας α = 0.02

Διατύπωση Υποθέσεων

H_0: μ ≤ 100

H_1: μ > 100

Ο έλεγχος που πρέπει να διεξαχθεί είναι μονοκατάληκτος προς τα πάνω

Άρα η μεταβλητή $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ στο άνω όριο έχει οριακή τιμή:

$$z_{(1-\alpha)} = z_{(1-0,02)} = Z_{0.98} = 2,05$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{121}}{\sqrt{50}} = \frac{11}{7.071} = 1.555649$$

Όταν έχουμε σφάλμα τύπου || σημαίνει ότι έχουμε αποδεχτεί την $H0$ χωρίς να ισχύει. Επομένως , η μεταβλητή Z θα πρέπει να βρίσκεται εντός της ζώνης αποδοχής.

Δηλαδή, πρέπει : $\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < z_{(1-\alpha)} \Rightarrow \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < 2,05 \Rightarrow \frac{\bar{X}-100}{1,55} * 1,55 < 2,05 * 1,55 \Rightarrow \bar{X} - 100 < 3,1775 \Rightarrow$

$$\bar{X} - 100 + 100 < 3,1775 + 100 \Rightarrow \bar{X} < 103,1775$$

Δίνεται μ=100.5

Σφάλμα Τύπου || Λανθασμένη μη απόρριψη, δηλαδή να μην απορρίψουμε λανθασμένα την H0

$$\beta = P\{\bar{X} < 103,1775\} = P\left\{\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{103,1775-100,5}{1.55}\right\} = P\left\{\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{2,6775}{1.55}\right\} = \Phi(1,73) = 0,9582$$

Άρα, η πιθανότητα σφάλματος τύπου || σε επίπεδο σημαντικότητας α=0,02 είναι Β=0,9582.

Συγκρίνετε τα αποτελέσματα των ερωτημάτων 1 και 2 και σχολιάστε τι προκύπτει από τη σύγκριση.

Μελετώντας τα δύο ερωτήματα, παρατηρούμε ότι η σχέση ανάμεσα στο (επίπεδο σημαντικότητας) α και την πιθανότητα σφάλματος τύπου || είναι αντιστρόφως ανάλογη. Με άλλα λόγια, όσο μειώνεται το επίπεδο σημαντικότητας, αυξάνεται η πιθανότητα σφάλματος τύπου ||.

Θέμα 5.

Σε μία προσπάθεια αξιολόγησης της ποιότητας ενός νέου λιπάσματος, επιλέξαμε 10 χωράφια τα οποία χωρίσαμε ακριβώς στη μέση. Στο ένα μισό χρησιμοποιήσαμε το παλιό λίπασμα και στο άλλο μισό χρησιμοποιήσαμε το νέο λίπασμα. Κατόπιν μετρήσαμε σε τόνους την σοδειά η οποία μαζεύτηκε από το κάθε κομμάτι των χωραφιών. Τα αποτελέσματα παρουσιάζονται στο Worksheet Fertilizer.

Α. Σε επίπεδο σημαντικότητας α = 10%, μπορούμε να ισχυριστούμε πως το νέο λίπασμα είναι αποδοτικότερο από το παλιό (με χρήση της μεθόδου των ανεξάρτητων δειγμάτων);

Β. Σε επίπεδο σημαντικότητας α = 10%, μπορούμε να ισχυριστούμε πως το νέο λίπασμα είναι αποδοτικότερο από το παλιό (με χρήση της μεθόδου των εξαρτημένων δειγμάτων);

Γ. Αν η αντιστοίχιση των αποτελεσμάτων των τμημάτων του ίδιου χωραφιού είχε γίνει με λανθασμένο τρόπο, όπως φαίνεται στο Worksheet Fertilizer, ποιο θα ήταν το αποτέλεσμα του ελέγχου μέσω της μεθόδου των εξαρτημένων δειγμάτων (ερώτημα Β). Σχολιάστε το αποτέλεσμα.

Α. . Σε επίπεδο σημαντικότητας α = 10%, μπορούμε να ισχυριστούμε πως το νέο λίπασμα είναι αποδοτικότερο από το παλιό (με χρήση της μεθόδου των ανεξάρτητων δειγμάτων);

Θέλουμε να δείξουμε ότι το νέο λίπασμα είναι αποδοτικότερο από το παλιό άρα θα έχουμε ένα μονοκατάληκτο έλεγχο προς τα πάνω όπου

$$H_0: \mu_y \leq \mu_x$$

$$H_1: \mu_y > \mu_x$$

Χωράφι	Τόνοι με Παλιό Λίπασμα Χ	Τόνοι με Νέο Λίπασμα Υ
1	18,828	19,776
2	20,064	21,912
3	20,316	24,684
4	21,252	24,972
5	24,108	25,224
6	24,24	25,932
7	24,48	26,352
8	24,72	27,132
9	28,632	27,168
10	29,22	33,804
Μέση Τιμή	23,586	25,6956
Διακύμανση	12,352552	13,5100816
τυπ. αποκλ.	3,51	3,68

Για τα δείγματα έχω:

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X\, i = 23{,}586$$

$$\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y\, i = 25{,}6956$$

Για τις δειγματικές διακυμάνσεις έχουμε:

$s_X^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$= 12,352552

$$s_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2 \ = 13{,}51$$

Οι τυπικές αποκλίσεις είναι:

$$s_X = \sqrt{s_X^2} = 3{,}51$$

$$s_Y = \sqrt{s_Y^2} = 3.68$$

Παρατηρώ ότι η ποσοστιαία διαφορά των τυπικών αποκλίσεων των δύο πληθυσμών είναι μικρή

$(\frac{s_Y - s_X}{s_X}100 = 4{,}58\%)$.

Άρα μπορώ να χρησιμοποιήσω την μέθοδο ανεξάρτητων δειγμάτων θεωρώντας s_X=s_Y. Επίσης το δείγμα είναι μικρό (n=10) άρα θα χρησιμοποιηθεί η κατανομή t-students.

Η κοινή διακύμανση θα είναι: $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = 12{,}93$

Υπολογισμός δειγματικής τυπικής απόκλισης: $s_{\bar{Y}-\bar{X}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = 1{,}6$

Το κριτήριο ελέγχου t είναι $t_v = \frac{\bar{Y}-\bar{X}-\theta}{s_{\bar{Y}-\bar{X}}}$ όπου θ από την υπόθεσή μας είναι 0 και ν οι βαθμοί ελευθερίας όπου ισχύει

ν= $n_1 + n_2 - 2$= 10 + 10 - 2 = 18

Άρα $t_{18} = 1{,}311$

Εφόσον είναι μονοκατάληκτος έλεγχος προς τα πάνω πρέπει να βρω την άνω οριακή τιμή η οποία από τους πίνακες t-students σε επίπεδο σημαντικότητας α=0,1 είναι $t_{1-a,\ 18} = 1{,}33$

Επειδή $t < t_{1-a,\ 18}$ δεν μπορούμε να απορρίψουμε την H_0 σε επίπεδο σημαντικότητας 10%.

Άρα δεν μπορούμε να υποστηρίξουμε ότι το νέο λίπασμα ήταν καλύτερο από το παλιό.

Β. Σε επίπεδο σημαντικότητας α = 10%, μπορούμε να ισχυριστούμε πως το νέο λίπασμα είναι αποδοτικότερο από το παλιό (με χρήση της μεθόδου των εξαρτημένων δειγμάτων);

Χωράφι	Τόνοι με Παλιό Λίπασμα Χ	Τόνοι με Νέο Λίπασμα Υ	D=X-Y
1	18,828	19,776	-0,948
2	20,064	21,912	-1,848
3	20,316	24,684	-4,368
4	21,252	24,972	-3,72
5	24,108	25,224	-1,116
6	24,24	25,932	-1,692
7	24,48	26,352	-1,872
8	24,72	27,132	-2,412
9	28,632	27,168	1,464
10	29,22	33,804	-4,584

Με την μέθοδο των εξαρτημένων δειγμάτων, θεωρώ την D=X-Y

$$H_0: \mu_D \geq 0$$

$$H_1: \mu_D < 0$$

Έχω ότι $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D\,i = -2{,}1$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 = 3{,}2571$$

$$s_D = \sqrt{S_D^2} = 1{,}8$$

Κριτήριο $t = \frac{\overline{D}}{s_D/\sqrt{n}} = -3{,}69$

Έχουμε μονοκατάληκτο έλεγχο προς τα κάτω άρα θα βρω την κάτω κρίσιμη τιμή από πίνακες t-students με βαθμούς ελευθερίας ν=10 και σε επίπεδο σημαντικότητας α= 0,1

$t_{0,1,\ 10=-1,372}$

Επειδή t < $t_{0,1,\ 10}$ απορρίπτεται η H_0, επομένως τα καινούργια λιπάσματα είναι καλύτερα απ' τα παλιά.

Γ. Αν η αντιστοίχιση των αποτελεσμάτων των τμημάτων του ίδιου χωραφιού είχε γίνει με λανθασμένο τρόπο, όπως φαίνεται στο Worksheet Fertilizer, ποιο θα ήταν το αποτέλεσμα του ελέγχου μέσω της μεθόδου των εξαρτημένων δειγμάτων (ερώτημα Β). Σχολιάστε το αποτέλεσμα.

Χωράφι	Τόνοι με Παλιό Λίπασμα	Τόνοι με Νέο Λίπασμα	D=X-Y
1	28,632	27,168	1,464
2	29,22	21,912	7,308
3	24,24	25,224	-0,984
4	21,252	24,972	-3,72
5	20,064	19,776	0,288
6	24,108	27,132	-3,024
7	20,316	26,352	-6,036
8	18,828	33,804	-14,976
9	24,72	25,932	-1,212
10	24,48	24,684	-0,204

Με την μέθοδο των εξαρτημένων δειγμάτων, θεωρώ πάλι την D=X-Y με μέγεθος δείγματος n=10 και τις εξής υποθέσεις:

$$H_0: \mu_D \geq 0$$

$$H_1: \mu_D < 0$$

Έχω ότι $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D\,i = -2{,}1$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 = 33{,}03$$

$$s_D = \sqrt{S_D^2} = 5{,}7475$$

Κριτήριο $t = \frac{\overline{D}}{s_D/\sqrt{n}} = -1{,}16$

Έχουμε μονοκατάληκτο έλεγχο προς τα κάτω άρα θα βρω την κάτω κρίσιμη τιμή από πίνακες t-students με βαθμούς ελευθερίας ν=10 και σε επίπεδο σημαντικότητας α= 0,1

$t_{0,1\,,\ 10=-1,372}$

Επειδή t > $t_{0,1\,,\ 10}$ **δεν** απορρίπτεται η H_0, επομένως **δεν** μπορούμε να υποστηρίξουμε ότι τα καινούργια λιπάσματα είναι καλύτερα.

Παρόλο που ο αριθμητής και στις δύο περιπτώσεις ήταν ο ίδιος ($\overline{D} = -2{,}1$), η διαφορά στο αποτέλεσμα οφείλεται στην μειωμένη τυπική απόκλιση που έχει η σωστή αντιστοίχιση για τα εξαρτημένα δείγματα, σε σχέση με την αντίστοιχη στην λανθασμένη αντιστοίχιση. Η οργάνωση των δεδομένων σε ομάδες μειώνει την διακύμανση εντός του δείγματος. Στα ανοργάνωτα δεδομένα παρατηρείται πολύ μεγάλη μεταβλητότητα ενώ στα οργανωμένα η διασπορά έχει μειωθεί (έχουμε εξουδετερώσει τη διασπορά που οφείλεται στα διαφορετικά βαθμολογικά γκρουπ), έτσι η τιμή φαίνεται σημαντική συγκριτικά με τη μεταβλητότητα που παρατηρείται στο δείγμα. Και στην περίπτωση μας στην λάθος αντιστοίχιση έχουμε διακύμανση 33,03 και τυπική απόκλιση 5,7475 ενώ στην σωστή αντιστοίχιση διακύμανση 3,2571 και τυπική απόκλιση 1,8. Η διαφορά είναι αρκετά μεγάλη και δείχνει ότι στην λανθασμένη αντιστοίχιση τα δείγματα δεν έχουν ομαδοποιηθεί με σωστό τρόπο και έχουν μεγάλη μεταβλητότητα που κάνει την διαφορά των μέσων των δύο δειγμάτων να φαίνεται μικρή.

Θέμα 6.

Το WorkSheet RentPrices περιέχει το κόστος ενοικίασης, καθώς και το εμβαδό σε τετραγωνικά μέτρα ενός δείγματος 15 διαμερισμάτων.

1. Προσδιορίστε την εξίσωση απλής γραμμικής παλινδρόμησης θεωρώντας εξαρτημένη μεταβλητή την τιμή ενοικίασης.

2. Να διεξαχθεί ο έλεγχος σημαντικότητας του συντελεστή κλίσης της γραμμής παλινδρόμησης σε στάθμη σημαντικότητας α = 5%. Ποιο είναι το p-value αυτού του ελέγχου;

3. Να διεξαχθεί ο έλεγχος σημαντικότητας του κριτηρίου F σε στάθμη σημαντικότητας α = 0.05. Προσδιορίστε το p-value με χρήση κάποιο στατιστικού πακέτου.

4. Σε επίπεδο εμπιστοσύνης 98%, προσδιορίστε τα διαστήματα εμπιστοσύνης για τη μέση τιμή της τιμής ενοικίασης ενός διαμερίσματος 60m^2 και ενός διαμερίσματος 95m^2 . Ποιο από τα δύο διαστήματα εμπιστοσύνης είναι μεγαλύτερο; Σχολιάστε το λόγο.

1. Προσδιορίστε την εξίσωση απλής γραμμικής παλινδρόμησης θεωρώντας εξαρτημένη μεταβλητή την τιμή ενοικίασης.

Έστω η πραγματική εξίσωση παλινδρόμησης Y = β0 + β1*X + ε

Εμείς μπορούμε να προσδιορίσουμε σύμφωνα με τη δειγματοληψία μας την $\hat{Y}$ = b0 + b1*X

Συμβολίζουμε με b0 και b1 τους συντελεστές παλινδρόμησης οι οποίοι προκύπτουν από το δείγμα του πληθυσμού και αποτελούν εκτιμήσεις των συντελεστών παλινδρόμησης του πληθυσμού β0 και β1 αντίστοιχα.

Ακολουθώντας τη μέθοδο των ελαχίστων τετραγώνων καταλήγουμε στους εξής τύπους:

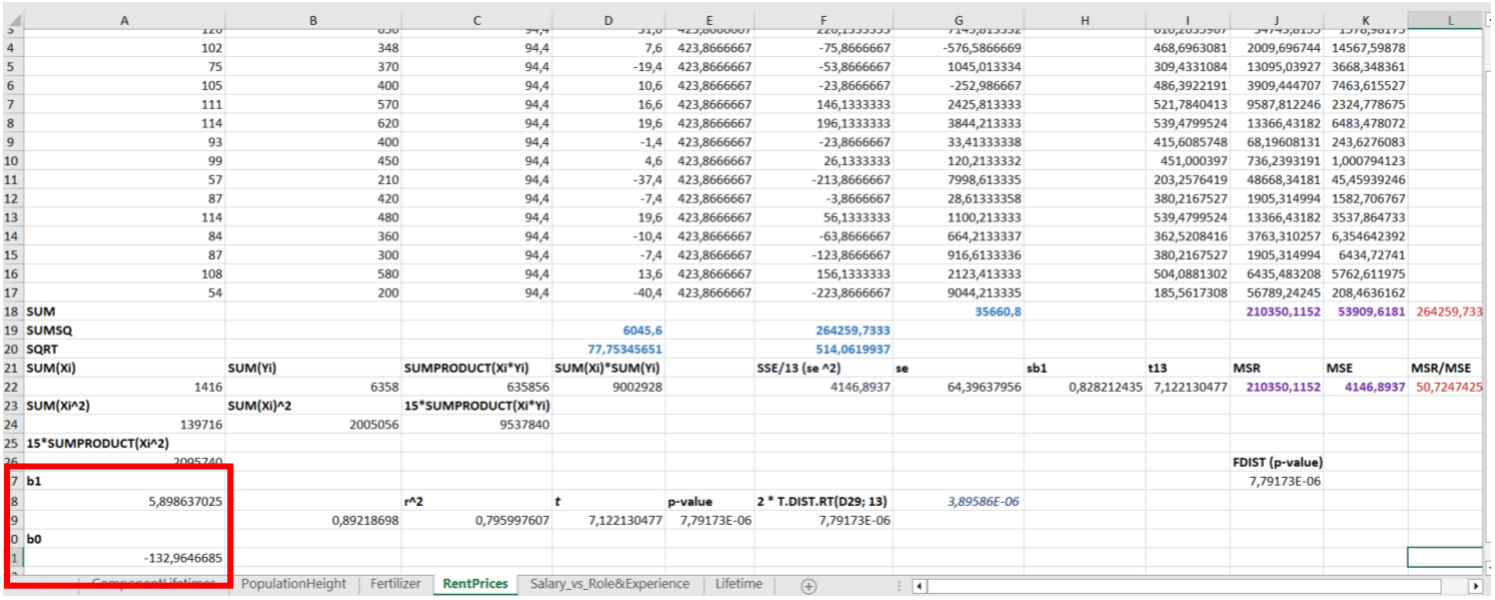
$$\boldsymbol{b1} = \frac{15 * \sum_{i=1}^{15}(Xi * Yi) - \sum_{i=1}^{15} Xi * \sum_{i=1}^{15} Yi}{15 * (\sum_{i=1}^{15} Xi^2) - \left(\sum_{i=1}^{15} Xi\right)^2} = \frac{15 * 635856 - 1416 * 6358}{15 * 139716 - 1416^2} =$$

$$= \frac{9537840 - 9002928}{2095740 - 2005056} = \frac{534912}{90684} = \mathbf{5.898637025} \approx \mathbf{5.8986}$$

$$\boldsymbol{b0} = \frac{\sum_{i=1}^{15} Yi}{15} - b1 * \frac{\sum_{i=1}^{15} Xi}{15} = \frac{6358}{15} - 5.8986 * \frac{1416}{15} = 423.8666 - 5.8986 * 94.4$$

$$= \mathbf{-132.9646685} \approx \mathbf{-132.965}$$

Άρα $\hat{Y}$ = -132.965 + 5.8986 * X



	A	B	C	D	E	F	G	H	I	J	K	L
4	102	348	94,4	7,6	423,8666667	-75,8666667	-576,5866669		468,6963081	2009,696744	14567,59878	
5	75	370	94,4	-19,4	423,8666667	-53,8666667	1045,013334		309,4331084	13095,03927	3668,348361	
6	105	400	94,4	10,6	423,8666667	-23,8666667	-252,986667		486,3922191	3909,444707	7463,615527	
7	111	570	94,4	16,6	423,8666667	146,1333333	2425,813333		521,7840413	9587,812246	2324,778675	
8	114	620	94,4	19,6	423,8666667	196,1333333	3844,213333		539,4799524	13366,43182	6483,478072	
9	93	400	94,4	-1,4	423,8666667	-23,8666667	33,41333338		415,6085748	68,19608131	243,6276083	
10	99	450	94,4	4,6	423,8666667	26,1333333	120,2133332		451,000397	736,2393191	1,000794123	
11	57	210	94,4	-37,4	423,8666667	-213,8666667	7998,613335		203,2576419	48668,34181	45,45939246	
12	87	420	94,4	-7,4	423,8666667	-3,8666667	28,61333358		380,2167527	1905,314994	1582,706767	
13	114	480	94,4	19,6	423,8666667	56,1333333	1100,213333		539,4799524	13366,43182	3537,864733	
14	84	360	94,4	-10,4	423,8666667	-63,8666667	664,2133337		362,5208416	3763,310257	6,354642392	
15	87	300	94,4	-7,4	423,8666667	-123,8666667	916,6133336		380,2167527	1905,314994	6434,72741	
16	108	580	94,4	13,6	423,8666667	156,1333333	2123,413333		504,0881302	6435,483208	5762,611975	
17	54	200	94,4	-40,4	423,8666667	-223,8666667	9044,213335		185,5617308	56789,24245	208,4636162	
18	**SUM**						35660,8			210350,1152	53909,6181	264259,733
19	**SUMSQ**			6045,6		264259,7333						
20	**SQRT**			77,75345651		514,0619937						
21	**SUM(Xi)**	**SUM(Yi)**	**SUMPRODUCT(Xi*Yi)**	**SUM(Xi)*SUM(Yi)**		**SSE/13 (se ^2)**	**se**	**sb1**	**t13**	**MSR**	**MSE**	**MSR/MSE**
22	1416	6358	635856	9002928		4146,8937	64,39637956	0,828212435	7,122130477	210350,1152	4146,8937	50,7247425
23	**SUM(Xi^2)**	**SUM(Xi)^2**	**15*SUMPRODUCT(Xi*Yi)**									
24	139716	2005056	9537840									
25	**15*SUMPRODUCT(Xi^2)**											
26	2095740									**FDIST (p-value)**		
27	**b1**									7,79173E-06		
28	5,898637025		**r^2**	***t***	**p-value**	**2 * T.DIST.RT(D29; 13)**	*3,89586E-06*					
29		0,89218698	0,795997607	7,122130477	7,79173E-06	7,79173E-06						
30	**b0**											
31	-132,9646685											

PopulationHeight | Fertilizer | RentPrices | Salary_vs_Role&Experience | Lifetime

2. Να διεξαχθεί ο έλεγχος σημαντικότητας του συντελεστή κλίσης της γραμμής παλινδρόμησης σε στάθμη σημαντικότητας α = 5%. Ποιο είναι το p-value αυτού του ελέγχου;

Έλεγχος Στατιστικής Σημαντικότητας του Συντελεστή Προσδιορισμού (Κλίση β_1)

Αρχικά υπολογίζουμε τον συντελεστή συσχέτισης r

$$r = \frac{\sum_{i=1}^{15}(Xi - \bar{X})\,(Yi - \bar{Y})}{\sqrt{\sum_{i=1}^{15}(Xi - \bar{X})^2} * \sqrt{\sum_{i=1}^{15}(Yi - \bar{Y})^2}} =$$

$$= \frac{35660.8}{77.75345651 * 514.0619937}$$

$$= \mathbf{0.89218698} \approx \mathbf{0.89}$$

$$\boldsymbol{r^2} = \mathbf{0,795998}$$

Θα ελέγξουμε αν η ποσότητα που προσδιορίζει το R^2 , δηλαδή το ποσοστό της μεταβλητότητας της Y που οφείλεται στις επιδράσεις της X είναι μηδενικό ή όχι

$$SSR = \sum_{i=1}^{15}(\hat{Y}\imath - \bar{Y})^2 = 210350{,}1152 > 0$$

$$SSE = \sum_{i=1}^{15}(Yi - \hat{Y}\imath)^2 = 53909{,}6181$$

$$SST = SSE + SSR = 264259{,}7333$$

$$\boldsymbol{R^2 = \frac{SSR}{SST} = \frac{210350{,}1152}{264259{,}7333} = 0{,}795998 = r^2}$$

Εφόσον το πηλίκο το οποίο υπολογίσαμε είναι ικανοποιητικά μεγάλο αυτό σημαίνει ότι η μεταβλητότητα της Y ερμηνεύεται σε μεγάλο βαθμό από τις επιδράσεις της X που μετριέται μέσω του SSR.

SSR > 0 άρα β1 ≠ 0 και η εξίσωση παλινδρόμησης εξηγεί τη διασπορά της y

Το b1 έχει μία κατανομή δειγματοληψίας η οποία περιγράφει πιθανοτικά πόσο “κοντά” είμαστε στο συντελεστή του πληθυσμού β1.

Το τυπικό σφάλμα (τυπική απόκλιση της κατανομής δειγματοληψίας) του b1 είναι το μέτρο του πόσο αδύναμη είναι η εκτίμηση που μας παρέχει ο συντελεστής.

Σύμφωνα με την υπόθεση της γραμμικής παλινδρόμησης, για δοθέν X, η κατανομή της Y είναι κανονική με μέση τιμή E[Y] και τυπική απόκλιση σ_Y

$$\sigma_{b1} = \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^{15}(Xi - \overline{X})^2}}$$

Επειδή η τυπική απόκλιση των σφαλμάτων στο υπόδειγμα της γραμμικής παλινδρόμησης είναι άγνωστη, θα χρησιμοποιήσουμε την εκτίμηση που μας παρέχει το δείγμα s_e

$$s_e{}^2 = \frac{SSE}{15-2} = \frac{53909{,}6181}{13} = 4146{,}8937$$

$$s_e = 64{,}39637956$$

$$\boldsymbol{s_{b1}} = \frac{s_e}{\sqrt{\sum_{i=1}^{15}(Xi - \overline{X})^2}} = \frac{64{,}39637956}{77{,}75345651} = 0{,}828212435 \approx \mathbf{0,83}$$

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Y											
2	τραγωνικά Μέτρα)	Τιμή Ενοικίασης (Μηνιαία)	AVERAGE X	Xi - X bar	AVERAGE Y	Yi - Y bar	(Xi - X bar)(Yi - Y bar)		Yi hat	SSR	SSE	SST	SSR/SST
3	126	650	94,4	31,6	423,8666667	226,1333333	7145,813332		610,2635967	34743,8155	1578,98175		0,795998
4	102	348	94,4	7,6	423,8666667	-75,8666667	-576,5866669		468,6963081	2009,696744	14567,59878		
5	75	370	94,4	-19,4	423,8666667	-53,8666667	1045,013334		309,4331084	13095,03927	3668,348361		
6	105	400	94,4	10,6	423,8666667	-23,8666667	-252,986667		486,3922191	3909,444707	7463,615527		
7	111	570	94,4	16,6	423,8666667	146,1333333	2425,813333		521,7840413	9587,812246	2324,778675		
8	114	620	94,4	19,6	423,8666667	196,1333333	3844,213333		539,4799524	13366,43182	6483,478072		
9	93	400	94,4	-1,4	423,8666667	-23,8666667	33,41333338		415,6085748	68,19608131	243,6276083		
10	99	450	94,4	4,6	423,8666667	26,1333333	120,2133332		451,000397	736,2393191	1,000794123		
11	57	210	94,4	-37,4	423,8666667	-213,8666667	7998,613335		203,2576419	48668,34181	45,45939246		
12	87	420	94,4	-7,4	423,8666667	-3,8666667	28,61333358		380,2167527	1905,314994	1582,706767		
13	114	480	94,4	19,6	423,8666667	56,1333333	1100,213333		539,4799524	13366,43182	3537,864733		
14	84	360	94,4	-10,4	423,8666667	-63,8666667	664,2133337		362,5208416	3763,310257	6,354642392		
15	87	300	94,4	-7,4	423,8666667	-123,8666667	916,6133336		380,2167527	1905,314994	6434,72741		
16	108	580	94,4	13,6	423,8666667	156,1333333	2123,413333		504,0881302	6435,483208	5762,611975		
17	54	200	94,4	-40,4	423,8666667	-223,8666667	9044,213335		185,5617308	56789,24245	208,4636162		
18							35660,8			210350,1152	53909,6181	264259,7333	
19				6045,6		264259,7333							
20				77,75345651		514,0619937							
21		SUM(Yi)	SUMPRODUCT(Xi*Yi)	SUM(Xi)*SUM(Yi)		SSE/13 (se ^2)	se	sb1	t13	MSR	MSE	MSR/MSE	
22	1416	6358	635856	9002928		4146,8937	64,39637956	0,828212435	7,122130477	210350,1152	4146,8937	50,72474253	
23		SUM(Xi)^2	15*SUMPRODUCT(Xi*Yi)										
24	139716	2005056	9537840										
25	)DUCT(Xi^2)												

Όταν η τυχαία μεταβλητή X ακολουθεί την κανονική κατανομή, τότε η τυχαία μεταβλητή $t_{n-2} = \frac{b1-\beta1}{s_{b1}}$ ακολουθεί την κατανομή t-student με n-2 βαθμούς ελευθερίας

Επομένως, το διάστημα εμπιστοσύνης για επίπεδο εμπιστοσύνης 1- α (επίπεδο σημαντικότητας α) του συντελεστή παλινδρόμησης του πληθυσμού β1 είναι:

$$b1 \pm s_{b1} * \left|t_{n-2,a/2}\right|$$

Διατύπωση Υποθέσεων

H0 : β1 = 0

H1 : β1 ≠ 0

$$t_{13} = \frac{b1-\beta1}{s_{b1}} = \frac{5.898637025 - 0}{0{,}828212435} = 7{,}122130477 \approx \mathbf{7,12}$$

$$t_{n-2,\frac{a}{2}} = -2.160 \quad \kappa\alpha\iota \quad t_{n-2,1-\frac{a}{2}} = 2.160$$

Επομένως

$$|t_{13}| > \left|t_{n-2,\frac{a}{2}}\right|$$

Δηλαδή η Υπόθεση β1 = 0 απορρίπτεται

Προσδιορισμός διαστήματος εμπιστοσύνης για το συντελεστή παλινδρόμησης του πληθυσμού β1

$$b1 \pm s_{b1} * \left|t_{n-2,\frac{a}{2}}\right| = 5.898637025 \pm 0{,}828212435 * |t_{13,0.025}|$$

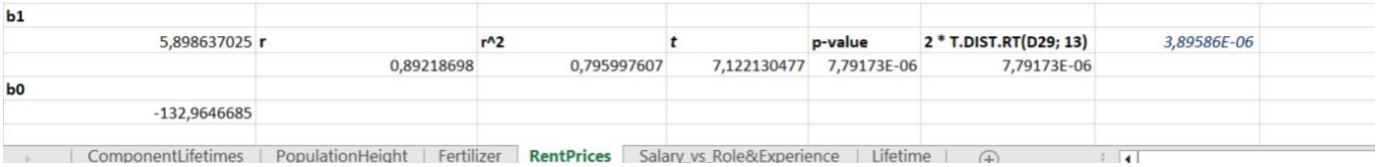
$$4{,}109698165 < \beta 1 < 7.687575885$$

Υπολογισμός **p-value** για δικατάληκτο έλεγχο

p = 2 * P { t_{13} > 7.12213 } = 2 * 0.00000389586 = 0.00000779173

Ισχύει p < α σε επίπεδο σημαντικότητας 0,05, *η Η0 απορρίπτεται*

Άρα $\beta 1 \neq 0$



b1						
5,898637025	**r**	**r^2**	***t***	**p-value**	**2 * T.DIST.RT(D29; 13)**	*3,89586E-06*
	0,89218698	0,795997607	7,122130477	7,79173E-06	7,79173E-06	
b0						
-132,9646685						

ComponentLifetimes | PopulationHeight | Fertilizer | **RentPrices** | Salary_vs_Role&Experience | Lifetime

3. Να διεξαχθεί ο έλεγχος σημαντικότητας του κριτηρίου F σε στάθμη σημαντικότητας α = 0.05. Προσδιορίστε το p-value με χρήση κάποιο στατιστικού πακέτου.

$k = 1$: Ένας βαθμός ελευθερίας λόγω της εκτίμησης του b1

$$MSR = \frac{SSR}{k} = SSR = 210350{,}1152$$

$$MSE = \frac{SSE}{n-k-1} = \frac{53909{,}6181}{13} = 4146{,}8937$$

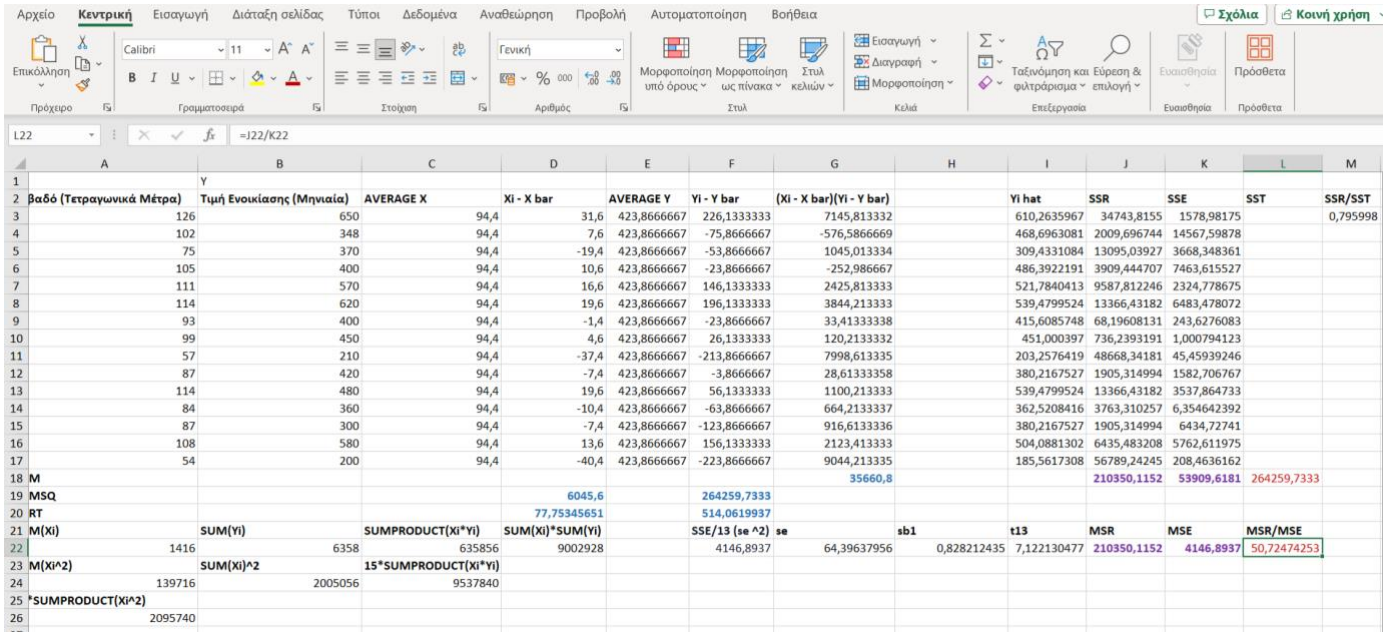
Βάση των MSR και MSE υπολογίζουμε το Κριτήριο Ελέγχου για τον έλεγχο της Η0 (β1=0)

$$\boldsymbol{F_{k,n-k-1} = \frac{MSR}{MSE}}$$

$$F_{1,13} = \frac{210350{,}1152}{4146{,}8937} = 50{,}72474253 \approx \mathbf{50{,}72}$$

$$F_{1,13,0.05} = 4{,}67$$

Ισχύει $F_{1,13} > F_{1,13,0.05}$, επομένως η Υπόθεση Η0 απορρίπτεται, δηλαδή $\beta 1 \neq 0$



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Y											
2	**βαδό (Τετραγωνικά Μέτρα)**	**Τιμή Ενοικίασης (Μηνιαία)**	**AVERAGE X**	**Xi - X bar**	**AVERAGE Y**	**Yi - Y bar**	**(Xi - X bar)(Yi - Y bar)**		**Yi hat**	**SSR**	**SSE**	**SST**	**SSR/SST**
3	126	650	94,4	31,6	423,8666667	226,1333333	7145,813332		610,2635967	34743,8155	1578,98175		0,795998
4	102	348	94,4	7,6	423,8666667	-75,8666667	-576,5866669		468,6963081	2009,696744	14567,59878		
5	75	370	94,4	-19,4	423,8666667	-53,8666667	1045,013334		309,4331084	13095,03927	3668,348361		
6	105	400	94,4	10,6	423,8666667	-23,8666667	-252,986667		486,3922191	3909,444707	7463,615527		
7	111	570	94,4	16,6	423,8666667	146,1333333	2425,813333		521,7840413	9587,812246	2324,778675		
8	114	620	94,4	19,6	423,8666667	196,1333333	3844,213333		539,4799524	13366,43182	6483,478072		
9	93	400	94,4	-1,4	423,8666667	-23,8666667	33,41333338		415,6085748	68,19608131	243,6276083		
10	99	450	94,4	4,6	423,8666667	26,1333333	120,2133332		451,000397	736,2393191	1,000794123		
11	57	210	94,4	-37,4	423,8666667	-213,8666667	7998,613335		203,2576419	48668,34181	45,45939246		
12	87	420	94,4	-7,4	423,8666667	-3,8666667	28,61333358		380,2167527	1905,314994	1582,706767		
13	114	480	94,4	19,6	423,8666667	56,1333333	1100,213333		539,4799524	13366,43182	3537,864733		
14	84	360	94,4	-10,4	423,8666667	-63,8666667	664,2133337		362,5208416	3763,310257	6,354642392		
15	87	300	94,4	-7,4	423,8666667	-123,8666667	916,6133336		380,2167527	1905,314994	6434,72741		
16	108	580	94,4	13,6	423,8666667	156,1333333	2123,413333		504,0881302	6435,483208	5762,611975		
17	54	200	94,4	-40,4	423,8666667	-223,8666667	9044,213335		185,5617308	56789,24245	208,4636162		
18	**M**						35660,8			210350,1152	53909,6181	264259,7333	
19	**MSQ**			6045,6		264259,7333							
20	**RT**			77,75345651		514,0619937							
21	**M(Xi)**	**SUM(Yi)**	**SUMPRODUCT(Xi*Yi)**	**SUM(Xi)*SUM(Yi)**		**SSE/13 (se ^2)**	**se**	**sb1**	**t13**	**MSR**	**MSE**	**MSR/MSE**	
22	1416	6358	635856	9002928		4146,8937	64,39637956	0,828212435	7,122130477	210350,1152	4146,8937	50,72474253	
23	**M(Xi^2)**	**SUM(Xi)^2**	**15*SUMPRODUCT(Xi*Yi)**										
24	139716	2005056	9537840										
25	***SUMPRODUCT(Xi^2)**												
26	2095740												

Υπολογισμός p – value

Χρησιμοποιώντας στο excel τη συνάρτηση FDIST, με F = 50,72474253 και βαθμούς ελευθερίας 1 και 13 αντίστοιχα με μονοκατάληκτο έλεγχο, βρίσκουμε ότι

$$\boldsymbol{p} = \mathbf{2} * \boldsymbol{P}\{F_{1,13} > 50{,}72474253\} = 0.00000779173$$

Ισχύει p < α σε επίπεδο σημαντικότητας 0,05, *η Η0 απορρίπτεται*

Άρα $\beta 1 \neq 0$

	A	B	C	D	E	F	G	H	I	J	K	L
1	X	Y										
2	**Εμβαδό (Τετραγωνικά Μέτρα)**	**Τιμή Ενοικίασης (Μηνιαία)**	**AVERAGE X**	**Xi - X bar**	**AVERAGE Y**	**Yi - Y bar**	**(Xi - X bar)(Yi - Y bar)**		**Yi hat**	**SSR**	**SSE**	**SST**
3	126	650	94,4	31,6	423,8666667	226,1333333	7145,813332		610,2635967	34743,8155	1578,98175	
4	102	348	94,4	7,6	423,8666667	-75,8666667	-576,5866669		468,6963081	2009,696744	14567,59878	
5	75	370	94,4	-19,4	423,8666667	-53,8666667	1045,013334		309,4331084	13095,03927	3668,348361	
6	105	400	94,4	10,6	423,8666667	-23,8666667	-252,986667		486,3922191	3909,444707	7463,615527	
7	111	570	94,4	16,6	423,8666667	146,1333333	2425,813333		521,7840413	9587,812246	2324,778675	
8	114	620	94,4	19,6	423,8666667	196,1333333	3844,213333		539,4799524	13366,43182	6483,478072	
9	93	400	94,4	-1,4	423,8666667	-23,8666667	33,41333338		415,6085748	68,19608131	243,6276083	
10	99	450	94,4	4,6	423,8666667	26,1333333	120,2133332		451,000397	736,2393191	1,000794123	
11	57	210	94,4	-37,4	423,8666667	-213,8666667	7998,613335		203,2576419	48668,34181	45,45939246	
12	87	420	94,4	-7,4	423,8666667	-3,8666667	28,61333358		380,2167527	1905,314994	1582,706767	
13	114	480	94,4	19,6	423,8666667	56,1333333	1100,213333		539,4799524	13366,43182	3537,864733	
14	84	360	94,4	-10,4	423,8666667	-63,8666667	664,2133337		362,5208416	3763,310257	6,354642392	
15	87	300	94,4	-7,4	423,8666667	-123,8666667	916,6133336		380,2167527	1905,314994	6434,72741	
16	108	580	94,4	13,6	423,8666667	156,1333333	2123,413333		504,0881302	6435,483208	5762,611975	
17	54	200	94,4	-40,4	423,8666667	-223,8666667	9044,213335		185,5617308	56789,24245	208,4636162	
18	**SUM**						35660,8			210350,1152	53909,6181	264259,733
19	**SUMSQ**			6045,6		264259,7333						
20	**SQRT**			77,75345651		514,0619937						
21	**SUM(Xi)**	**SUM(Yi)**	**SUMPRODUCT(Xi*Yi)**	**SUM(Xi)*SUM(Yi)**		**SSE/13 (se ^2)**	**se**	**sb1**	**t13**	**MSR**	**MSE**	**MSR/MSE**
22	1416	6358	635856	9002928		4146,8937	64,39637956	0,828212435	7,122130477	210350,1152	4146,8937	50,7247425
23	**SUM(Xi^2)**	**SUM(Xi)^2**	**15*SUMPRODUCT(Xi*Yi)**									
24	139716	2005056	9537840									
25	**15*SUMPRODUCT(Xi^2)**											
26	2095740									**FDIST (p-value)**		
27	**b1**									7,79173E-06		
28	5,898637025	**r**	**r^2**	**t**	**p-value**	**2 * T.DIST.RT(D29; 13)**	*3,89586E-06*					
29		0,89218698	0,795997607	7,122130477	7,79173E-06	7,79173E-06						

ComponentLifetimes | PopulationHeight | Fertilizer | **RentPrices** | Salary_vs_Role&Experience | Lifetime

Έτοιμο | Προσβασιμότητα: Διερεύνηση

4.Σε επίπεδο εμπιστοσύνης 98%, προσδιορίστε τα διαστήματα εμπιστοσύνης για τη μέση τιμή της τιμής ενοικίασης ενός διαμερίσματος 60m^2 και ενός διαμερίσματος 95m^2 . Ποιο από τα δύο διαστήματα εμπιστοσύνης είναι μεγαλύτερο; Σχολιάστε το λόγο.

Βάσει της γραμμής παλινδρόμησης, η πρόβλεψη για $60m^2$ είναι:

$$\widehat{Y_0} = b0 + b1 * X_0 = 5.898637025 + -132.9646685 * 60 = 220{,}953553 \approx 220{,}953$$

$$\widehat{Y_0} \pm t_{n-2,a/2} * \sqrt{\frac{SSE}{n-2}} * \sqrt{\frac{1}{n} + \frac{(X0 - \bar{X})^2}{\sum_{i=1}^{n} Xi^2 - \frac{\left(\sum_{i=1}^{n} Xi\right)^2}{n}}}$$

$$\widehat{Y_0} \pm t_{n-2,a/2} * s_{\widehat{Y_0}}$$

Επομένως για $\alpha = 1 - 0.98 = 0.02$, έχουμε $t_{13,0.02/2} = t_{13,0.01} = 2{,}650$

$$\boldsymbol{s_{\widehat{Y_0}}} = \sqrt{\frac{53909{,}6181}{13}} * \sqrt{\frac{1}{15} + \frac{(60 - 94{,}4)^2}{139716 - \frac{2005056}{15}}} = 64{,}39637956 * 0{,}512255519 = \mathbf{32{,}98740082}$$

$$\approx \mathbf{32{,}987}$$

Άρα το διάστημα εμπιστοσύνης για τη μέση τιμή $E[Y_0]$ είναι:

$$E[Y_0] \in [220{,}953 - 2{,}650 * 32{,}987\ ,220{,}953 + 2{,}650 * 32{,}987]$$

$$133{,}536 < E[Y_0] < 308{,}369$$

	A	B	C	D	E	F	G
22	1416	6358	635856	9002928		4146,8937	64,39637956
23	SUM(Xi^2)	SUM(Xi)^2	15*SUMPRODUCT(Xi*Yi)				
24	139716	2005056	9537840				
25	15*SUMPRODUCT(Xi^2)						
26	2095740						
27	b1						
28	5,898637025	r	r^2	t	p-value	2 * T.DIST.RT(D29; 13)	3,89586E-06
29		0,89218698	0,795997607	7,122130477	7,79173E-06	7,79173E-06	
30	b0						
31	-132,9646685						
32							
33	Y0 για X0= 60	Y0 για X0 = 95	SUM Xi^2	(SUM Xi) ^2	SQRT(SSE/n-2)	SQRT(1/13 + ((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15)))	S Y0 hat (result)
34	220,953553	427,4058489	139716	2005056	64,39637956	0,512255519	32,98740082
35							
36			(X0 - X bar)^2	((SUM Xi) ^2)/15			
37	t 13, 0.01		1183,36	133670,4			
38	2.650		(1/15)	((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15))			
39			0,066666667	0,19573905			
40							
41			SUM Xi^2	(SUM Xi) ^2	SQRT(SSE/n-2)	SQRT(1/13 + ((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15)))	S Y0 hat (result)
42			139716	2005056	64,39637956	0,258314177	16,6344978
43							
44			(X0 - X bar)^2	((SUM Xi) ^2)/15			
45			0,36	133670,4			
46			(1/15)	((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15))			
47			0,066666667	5,95474E-05			
48							
49							

Βάσει της γραμμής παλινδρόμησης, η πρόβλεψη για $95m^2$ είναι:

$$\widehat{Y_0} = b0 + b1 * X_0 = 5.898637025 + -132.9646685 * 95 = 427{,}4058489 \approx 427{,}405$$

$$\widehat{Y_0} \pm t_{n-2,a/2} * \sqrt{\frac{SSE}{n-2}} * \sqrt{\frac{1}{n} + \frac{(X0 - \bar{X})^2}{\sum_{i=1}^{n} Xi^2 - \frac{\left(\sum_{i=1}^{n} Xi\right)^2}{n}}}$$

$$\widehat{Y_0} \pm t_{n-2,a/2} * s_{\widehat{Y_0}}$$

Επομένως για $\alpha = 1 - 0.98 = 0.02$, έχουμε $t_{13,0.02/2} = t_{13,0.01} = 2{,}650$

$$s_{\widehat{Y_0}} = \sqrt{\frac{53909{,}6181}{13}} * \sqrt{\frac{1}{15} + \frac{(95 - 94{,}4)^2}{139716 - \frac{2005056}{15}}} = 64{,}39637956 * 0{,}258314177 = \mathbf{16{,}6344978} \approx \mathbf{16{,}634}$$

Άρα το διάστημα εμπιστοσύνης για τη μέση τιμή $E[Y_0]$ είναι:

$$E[Y_0] \in [220{,}953 - 2{,}650 * 16{,}634\ ,220{,}953 + 2{,}650 * 16{,}634]$$

$$176{,}871 < E[Y_0] < 265{,}034$$

	A	B	C	D	E	F	G
22	1416	6358	635856	9002928		4146,8937	64,39637956
23	SUM(Xi^2)	SUM(Xi)^2	15*SUMPRODUCT(Xi*Yi)				
24	139716	2005056	9537840				
25	15*SUMPRODUCT(Xi^2)						
26	2095740						
27	b1						
28	5,898637025	r	r^2	t	p-value	2 * T.DIST.RT(D29; 13)	3,89586E-06
29		0,89218698	0,795997607	7,122130477	7,79173E-06	7,79173E-06	
30	b0						
31	-132,9646685						
32							
33	Y0 για X0= 60	Y0 για X0 = 95	SUM Xi^2	(SUM Xi) ^2	SQRT(SSE/n-2)	SQRT(1/13 + ((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15)))	S Y0 hat (result)
34	220,953553	427,4058489	139716	2005056	64,39637956	0,512255519	32,98740082
35							
36			(X0 - X bar)^2	((SUM Xi) ^2)/15			
37	t 13, 0.01		1183,36	133670,4			
38	2.650		(1/15)	((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15))			
39			0,066666667	0,19573905			
40							
41			SUM Xi^2	(SUM Xi) ^2	SQRT(SSE/n-2)	SQRT(1/13 + ((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15)))	S Y0 hat (result)
42			139716	2005056	64,39637956	0,258314177	16,6344978
43							
44			(X0 - X bar)^2	((SUM Xi) ^2)/15			
45			0,36	133670,4			
46			(1/15)	((X0 - X bar)^2/(SUM Xi^2 -(((SUM Xi) ^2)/15))			
47			0,066666667	5,95474E-05			
48							
49							

Παρατηρούμε ότι το διάστημα εμπιστοσύνης για τα 95 m^2 είναι μικρότερο από αυτό των 60 m^2.

Αυτό οφείλεται κατά κύριο λόγω στην απόκλιση των τιμών ενοικίασης στα 2 διαφορετικά μεγέθη διαμερισμάτων.

Στα 60 m^2 υπάρχει μεγαλύτερο τυπικό σφάλμα $\boldsymbol{s_{\widehat{Y_0}} \approx 32{,}987}$ ενώ στα 95 m^2 μικρότερο $\boldsymbol{s_{\widehat{Y_0}} \approx 16{,}634}$

Έτσι οδηγούμαστε σε ένα ευρύτερο και ένα μικρότερο διάαστημα εμπιστοσύνης αντίστοιχα, λόγω της απόκλισης των τιμών στην αγορά.

Θέμα 7.

Στο Salary_vs_Role&Experience παρουσιάζεται το μέσο μεικτό μηνιαίο εισόδημα ενός δείγματος 44 senior εργαζομένων με έδρα την Γερμανία. Για κάθε άτομο του δείγματος σημειώνεται ο ρόλος του, καθώς και τα χρόνια συνολικής προϋπηρεσίας στο ρόλο αυτό.

1. Υπολογίστε τη γραμμή παλινδρόμησης με εξαρτημένη μεταβλητή το μεικτό μηνιαίο εισόδημα και ανεξάρτητη την προϋπηρεσία των εργαζομένων και ελέγξτε τη σημαντικότητα του κριτηρίου F σε α = 0.05.

2. Υποθέτοντας πως η προϋπηρεσία και ο επαγγελματικός ρόλος δεν έχουν αλληλεπίδραση (interaction), κατασκευάστε το υπόδειγμα πολλαπλής παλινδρόμησης θεωρώντας εξαρτημένη μεταβλητή το μεικτό μηνιαίο μισθό και ανεξάρτητες την προϋπηρεσία και τον επαγγελματικό ρόλο. Ελέγξτε τη σημαντικότητα του κριτηρίου F σε α = 0.05, καθώς και των διάφορων συντελεστών της πολλαπλής παλινδρόμησης.

3. Τι συμπεράσματα προκύπτουν από τη σχέση των δύο ελέγχων σημαντικότητας των ερωτημάτων 1 & 2.

4. Σχεδιάστε σε ένα διάγραμμα τις τρεις συναρτήσεις οι οποίες προβλέπουν τη μέση τιμή μεικτού μηνιαίου μισθού ως συνάρτηση της προϋπηρεσίας για κάθε έναν από τους επαγγελματικούς ρόλους (μία γραμμή για κάθε ρόλο).

1. Υπολογίστε τη γραμμή παλινδρόμησης με εξαρτημένη μεταβλητή το μεικτό μηνιαίο εισόδημα και ανεξάρτητη την προϋπηρεσία των εργαζομένων και ελέγξτε τη σημαντικότητα του κριτηρίου F σε α = 0.05.

Έστω η πραγματική εξίσωση παλινδρόμησης Y = β0 + β1*X + ε

Εμείς μπορούμε να προσδιορίσουμε σύμφωνα με τη δειγματοληψία μας την Ŷ = b0 + b1*X

Συμβολίζουμε με b0 και b1 τους συντελεστές παλινδρόμησης οι οποίοι προκύπτουν από το δείγμα του πληθυσμού και αποτελούν εκτιμήσεις των συντελεστών παλινδρόμησης του πληθυσμού β0 και β1 αντίστοιχα.

Ακολουθώντας τη μέθοδο των ελαχίστων τετραγώνων καταλήγουμε στους εξής τύπους:

$$\boldsymbol{b1} = \frac{44 * \sum_{i=1}^{44}(Xi * Yi) - \sum_{i=1}^{44} Xi * \sum_{i=1}^{44} Yi}{44 * (\sum_{i=1}^{44} Xi^2) - \left(\sum_{i=1}^{44} Xi\right)^2} = \frac{44 * 3608195{,}228 - 158302458{,}5}{44 * 6361{,}66 - 261529{,}96} =$$

$$= \frac{158760590 - 158302458{,}5}{279913{,}04 - 261529{,}96} = \frac{458131{,}5}{18383{,}1} = \boldsymbol{24{,}92136769 \approx 24{,}921}$$

$$\boldsymbol{b0} = \frac{\sum_{i=1}^{44} Yi}{44} - b1 * \frac{\sum_{i=1}^{44} Xi}{44} = \frac{309547{,}24}{44} - 24{,}92136769 * \frac{511{,}40}{44} = 7035{,}2 - 24{,}92136769 * 11{,}622727$$
$$= \boldsymbol{6745{,}510286 \approx 6745{,}510}$$

Άρα Ŷ = 6745,510+ 24,921 * X1

Με τη χρήση του στατιστικού πακέτου IBM SPSS παίρνουμε τον παρακάτω πίνακα

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6745,510	918,580		7,343	,000
	Προϋπηρεσία	24,921	76,394	,050	,326	,746

a. Dependent Variable: Μισθός

	A	B	C	D	E	F	G	H	I
1		X1				**Ερώτημα 1**			
2	**Εργαζόμενος**	**Προυπηρεσία**	**Ρόλος**	**Μεικτός Μηνιαίος Μισθός (Ευρώ)**		**SUM(Xi)**	**SUM(Yi)**	**SUMPRODUCT(Xi*Yi)**	**SUM(Xi)*SUM(Yi)**
3	1	11,00	Business	6908,40		511,40	309547,24	3608195,228	158302458,5
4	2	11,30	Business	7896,40					
5	3	11,40	Business	8215,60		**SUM(Xi^2)**	**SUM(Xi)^2**	**44*SUMPRODUCT(Xi*Yi)**	**44*SUMPRODUCT(Xi^2)**
6	4	13,40	Business	9235,52		6361,66	261529,96	158760590	279913,04
7	5	9,40	Business	7209,36					
8	6	14,20	Business	9901,28		458131,496		**b1**	
9	7	11,50	Business	6849,12		18383,1		24,92136769	
10	8	11,50	Business	7136,40					
11	9	11,50	Business	9040,20		**309547,24/44**	**511,40/44**	**b0**	
12	10	14,20	Business	10912,84		7035,2	11,622727	6745,510286	
13	11	9,60	Business	7608,36					
14	12	13,90	Business	9348,76					
15	13	8,90	Business	7061,16					
16	14	9,40	Business	7377,32					
17	15	9,10	Business	6185,64					
18	16	12,70	Data Scientist	6791,36					
19	17	16,40	Data Scientist	6425,80					
20	18	16,00	Data Scientist	6371,84					
21	19	15,50	Data Scientist	5931,80					
22	20	17,90	Data Scientist	6913,72					
23	21	14,60	Data Scientist	5982,72					
24	22	13,40	Data Scientist	8232,32					
25	23	15,00	Data Scientist	5648,32					
26	24	16,00	Data Scientist	7364,40					
27	25	12,10	Data Scientist	2800,60					
28	26	16,80	Data Scientist	5435,52					
29	27	15.80	Data Scientist	6543.60					

$$SSR = \sum_{i=1}^{44} (\hat{Y}i - \bar{Y})^2 = 259478{,}261$$

$$SSE = \sum_{i=1}^{44} (Yi - \hat{Y}i)^2 = 102407308{,}215$$

$k = 1$: Ένας βαθμός ελευθερίας λόγω της εκτίμησης του b1

$$MSR = \frac{SSR}{k} = SSR = 259478{,}261$$

$$MSE = \frac{SSE}{n-k-1} = \frac{102407308{,}215}{42} = 2438269{,}243$$

Διατύπωση Υποθέσεων

H0: $\beta 1 = 0$

H1: $\beta 1 \neq 0$

Βάση των MSR και MSE υπολογίζουμε το Κριτήριο Ελέγχου για τον έλεγχο της H0 (β1=0)

$$\boldsymbol{F_{k,n-k-1} = \frac{MSR}{MSE}}$$

$$F_{1,42} = \frac{259478{,}261}{2438269{,}243} = 0{,}106419114 \approx \mathbf{0,106}$$

Με τη χρήση του στατιστικού πακέτου IBM SPSS παίρνουμε τον παρακάτω πίνακα

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	259483,261	1	259483,261	,106	,746[b]
	Residual	102407308,215	42	2438269,243		
	Total	102666791,476	43			

a. Dependent Variable: Μισθός

b. Predictors: (Constant), Προϋπηρεσία

Άρα η τιμή του κριτηρίου F είναι όντως 0,106 και το sig είναι 0,746 > 0,05 άρα δεχόμαστε την υπόθεση H0

με την εξίσωση παλινδρόμησης να μην εξηγεί καθόλου τη διασπορά της Y ($\beta 1 = 0$ στον πληθυσμό)

2. Υποθέτοντας πως η προϋπηρεσία και ο επαγγελματικός ρόλος δεν έχουν αλληλεπίδραση (interaction), κατασκευάστε το υπόδειγμα πολλαπλής παλινδρόμησης θεωρώντας εξαρτημένη μεταβλητή το μεικτό μηνιαίο μισθό και ανεξάρτητες την προϋπηρεσία και τον επαγγελματικό ρόλο. Ελέγξτε τη σημαντικότητα του κριτηρίου F σε α = 0.05, καθώς και των διάφορων συντελεστών της πολλαπλής παλινδρόμησης.

Θα εφαρμόσουμε το μοντέλο της πολλαπλής παλινδρόμησης. Η άσκηση θα λυθεί με τη βοήθεια του στατιστικού πακέτου IBM SPSS Statistics.

Πρώτα απ' όλα χρειάζεται οι μεταβλητές που θα συμμετέχουν να είναι όλες scale. Η προϋπηρεσία είναι scale αλλά ο επαγγελματικός ρόλος είναι nominal.

Συνεπώς δημιουργούμε k-1 Dummy ψευδομεταβλητές, δηλαδή εφόσον έχουμε 3 ρόλους (1= Business, 2=Data Scientist, 3=Software Engineer), θα δημιουργήσουμε 2 μεταβλητές τις Dummy1 και Dummy2.

Έτσι έχουμε:

	Dummy1	**Dummy2**
1=Business	0	0
2=Data Scientist	0	1
3=Software Engineer	1	0

Έπειτα ελέγχουμε την ανεξαρτησία των μεταβλητών, οι οποίες μας δίνεται ότι είναι ανεξάρτητες, άρα συνεχίζουμε εφαρμόζοντας στο στατιστικό πακέτο την πολλαπλή παλινδρόμηση, ελέγχοντας για τον αν υπάρχει ομοσκεδαστικότητα, συγγραμμικότητα και αυτοσυσχέτιση.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	7553,430	255,992		29,506	,000		
	Dummy1	-1520,246	438,437	-,472	-3,467	,001	1,000	1,000
2	(Constant)	2813,836	755,245		3,726	,001		
	Dummy1	-3871,576	479,259	-1,201	-8,078	,000	,424	2,356
	Προϋπηρεσία	476,754	73,723	,962	6,467	,000	,424	2,356

a. Dependent Variable: Μισθός

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Προϋπηρεσία	,962[b]	6,467	,000	,711	,424	2,356	,424
	Dummy2	.[b]	.	.	.	,000	.	,000
2	Dummy2	.[c]	.	.	.	,000	.	,000

a. Dependent Variable: Μισθός

b. Predictors in the Model: (Constant), Dummy1

c. Predictors in the Model: (Constant), Dummy1, Προϋπηρεσία

Πρώτα απ'όλα από τους παραπάνω πίνακες βλέπουμε ότι στην παλινδρόμηση συμμετέχουν τελικά μόνο η Προϋπηρεσία και η Dummy1 καθώς η Dummy2 φαίνεται να παρουσιάσει μερική συσχέτιση με την «προϋπηρεσία» κατά 71,1% με στατιστική σημαντικότητα . Συνεπώς η ανάλυση περιορίζεται στις 2 μεταβλητές «Προϋπηρεσία» και Dummy1.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,472[a]	,223	,204	1378,55974	
2	,784[b]	,615	,596	981,71236	2,027

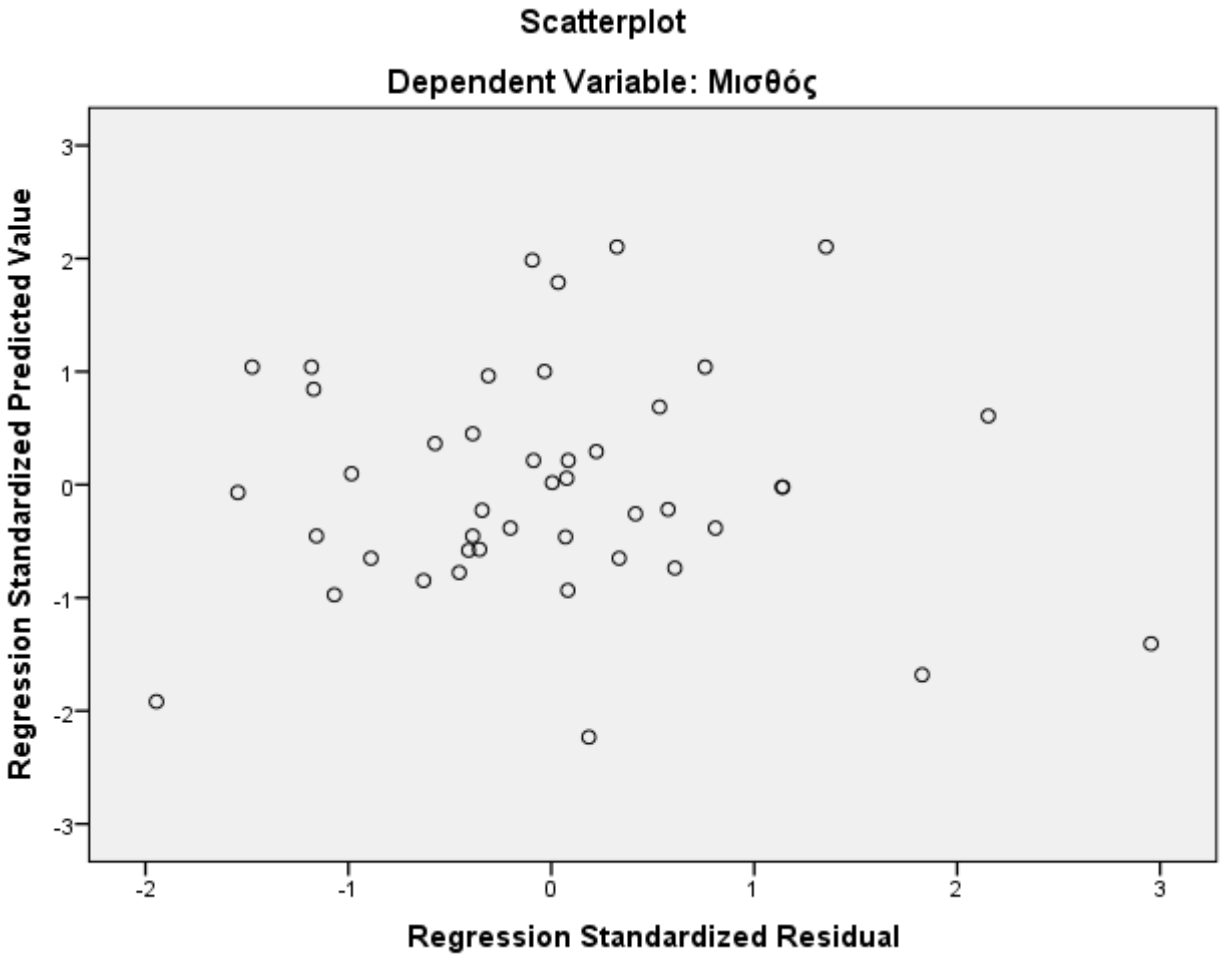
a. Predictors: (Constant), Dummy1

b. Predictors: (Constant), Dummy1, Προϋπηρεσία

c. Dependent Variable: Μισθός

Αντίθετα, σύμφωνα με τον παραπάνω πίνακα οι 2 μεταβλητές που συμμετέχουν τελικά στην παλινδρόμηση δεν παρουσιάζουν καθόλου αυτοσυσχέτιση αφού ο δείκτης "Durbin-Watson" βρίσκεται μεταξύ των επιτρεπτών ορίων 1,5 < 2,027 < 2,5 έχοντας μία πολύ καλή προσέγγιση στο 2.

Όσον αφορά στη συγγραμμικότητα, παρατηρείται μια πολύ μικρή στα όρια του αποδεκτού, καθώς το "Tolerance" και των δύο είναι 0.3 < 0,424 και "VIF" 2,356 < 3.33 από τον πίνακα "Coefficients" παραπάνω.



Τέλος από το δράγραμμα “Scatterplot”, βλέπουμε ότι τα κατάλοιπα είναι σχετικά ομοιόμορφα κατανεμημένα, οπότε υπάρχει ομοσκεδαστικότητα.

Από τον πίνακα λοιπόν “Coefficients”, παίρνουμε τις τιμές των “Beta” και προκύπτει η εξής εξίσωση παλινδρόμησης:

$\hat{Y} = \mathbf{2813,836} + \mathbf{476,754} * \mathbf{X1} - \mathbf{3871,756} * \boldsymbol{D1}$, όπου X1 η προϋπηρεσία και D1 η Dummy1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22848858,927	1	22848858,927	12,023	,001[b]
	Residual	79817932,549	42	1900426,965		
	Total	102666791,476	43			
2	Regression	63152665,948	2	31576332,974	32,764	,000[c]
	Residual	39514125,528	41	963759,159		
	Total	102666791,476	43			

a. Dependent Variable: Μισθός

b. Predictors: (Constant), Dummy1

c. Predictors: (Constant), Dummy1, Προϋπηρεσία

Λαμβάνοντας υπόψιν τη 2η γραμμή όπου έχουμε την εξήγηση των **SSR** = 63152665,948, **SSE**=39514125,528,

MSR=31576332,974 και **MSE** = 963759,159 και για τις 2 μεταβλητές. Συνεπώς το $F_{2,41} = 32{,}764$ όπου το $sig = 0 < 0.05$, άρα οι συντελεστές και κατά συνέπεια η εξίσωση παλινδρόμησης ισχύει και στον πληθυσμό.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,472[a]	,223	,204	1378,55974	
2	,784[b]	,615	,596	981,71236	2,027

a. Predictors: (Constant), Dummy1

b. Predictors: (Constant), Dummy1, Προϋπηρεσία

c. Dependent Variable: Μισθός

Επιπρόσθετα, σύμφωνα με τον πίνακα "Model Summary" μπορούμε να καταλάβουμε και το κατά πόσο οι 2 μεταβλητές εξηγούν τη συμπεριφορά της εξαρτημένης Y. Αφού λοιπόν έχουμε επιβεβαιώσει παραπάνω την υπόθεση **H1:** $\boldsymbol{R^2 \neq 0}$ μέσω ANOVA, καταλαβαίνουμε ότι η Dummy1 εξηγεί το 22,3% της διακύμανσης της εξαρτημένης ενώ και η Dummy1 και η «προϋπηρεσία» το 61,5% της διακύμανσης της εξαρτημένης, δηλαδή του μισθού.

3. Τι συμπεράσματα προκύπτουν από τη σχέση των δύο ελέγχων σημαντικότητας των ερωτημάτων 1 & 2.

Αρχικά το πρώτο μοντέλο απλής γραμμικής παλινδρόμησης που λαμβάνει υπόψιν μόνο την προϋπηρεσία κρίνεται ανεπαρκές καθώς η διακύμανση των τιμών του μισθού δεν μπορεί να εξηγηθεί σε σημαντικό βαθμό μόνο από την προϋπηρεσία.

Αντίθετα λαμβάνοντας υπόψιν και τον ρόλο των εργαζομένων, έχουμε μια πιο αντικειμενική εικόνα. Δημιουργώντας 2 ψευδομεταβλητές καταφέραμε να κωδικοποιήσουμε τους ρόλους τους και να δούμε πώς ο καθένας επιδρά σε συνδυασμό με την προϋπηρεσία στον μισθό. Ωστόσο χρειάστηκε να απορρίψουμε τη μία εκ των δύο ψευδομεταβλητών αφού αποδείχθηκε ότι δε συνεισέφερε επαρκώς στο μοντέλο και μη πληρώντας ορισμένες προϋποθέσεις. Έτσι καταλήξαμε στο μοντέλο της πολλαπλής παλινδρόμησης $\hat{Y} = \mathbf{2813,836} + \mathbf{476,754} * \mathbf{X1} - \mathbf{3871,756} * \boldsymbol{D1}$ το οποίο εξηγεί και το 61,5% της διακύμανσης των μισθών, δηλαδή ένα αρκετά ικανοποιητικό ποσοστό για τη διοικητική επιστήμη.

4. Σχεδιάστε σε ένα διάγραμμα τις τρεις συναρτήσεις οι οποίες προβλέπουν τη μέση τιμή μεικτού μηνιαίου μισθού ως συνάρτηση της προϋπηρεσίας για κάθε έναν από τους επαγγελματικούς ρόλους (μία γραμμή για κάθε ρόλο).

Ουσιαστικά εφαρμόζουμε τρεις επιμέρους απλές γραμμικές παλινδρομήσεις για κάθε ένα ρόλο, με μοναδική ανεξάρτητη μεταβλητή την προϋπηρεσία.

Για τον ρόλο "Business" σχηματίζεται ένα δείγμα από 15 άτομα, και μέσω του SPSS προκύπτει η εξής γραμμική εξίσωση

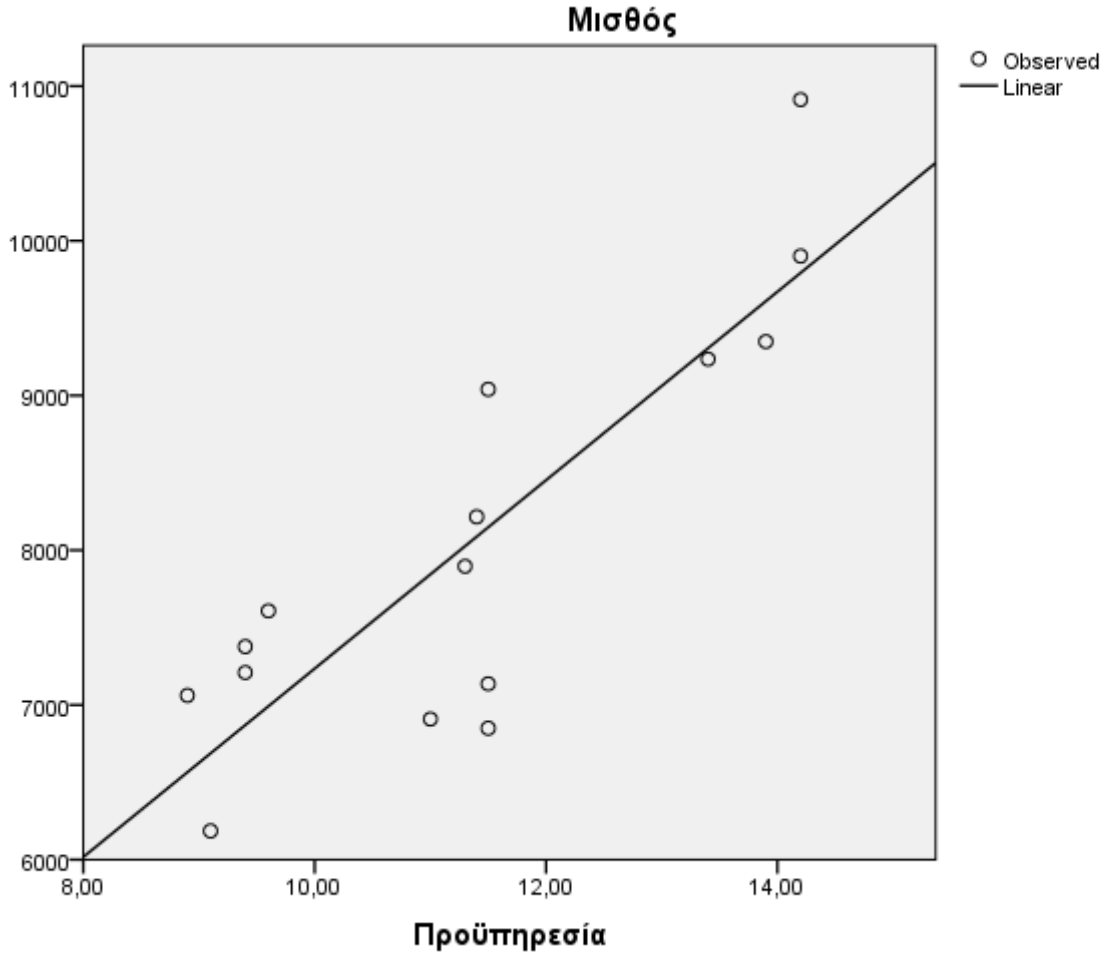
$$\hat{\boldsymbol{Y}} = \mathbf{1145,181} + \mathbf{608,976} * \boldsymbol{X}$$

Model Summary and Parameter Estimates

Dependent Variable: Μισθός

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	,721	33,628	1	13	,000	1145,181	608,976

The independent variable is Προϋπηρεσία.



Για τον ρόλο “Data Scientist” σχηματίζεται ένα δείγμα από 15 άτομα, και μέσω του SPSS προκύπτει η εξής γραμμική εξίσωση

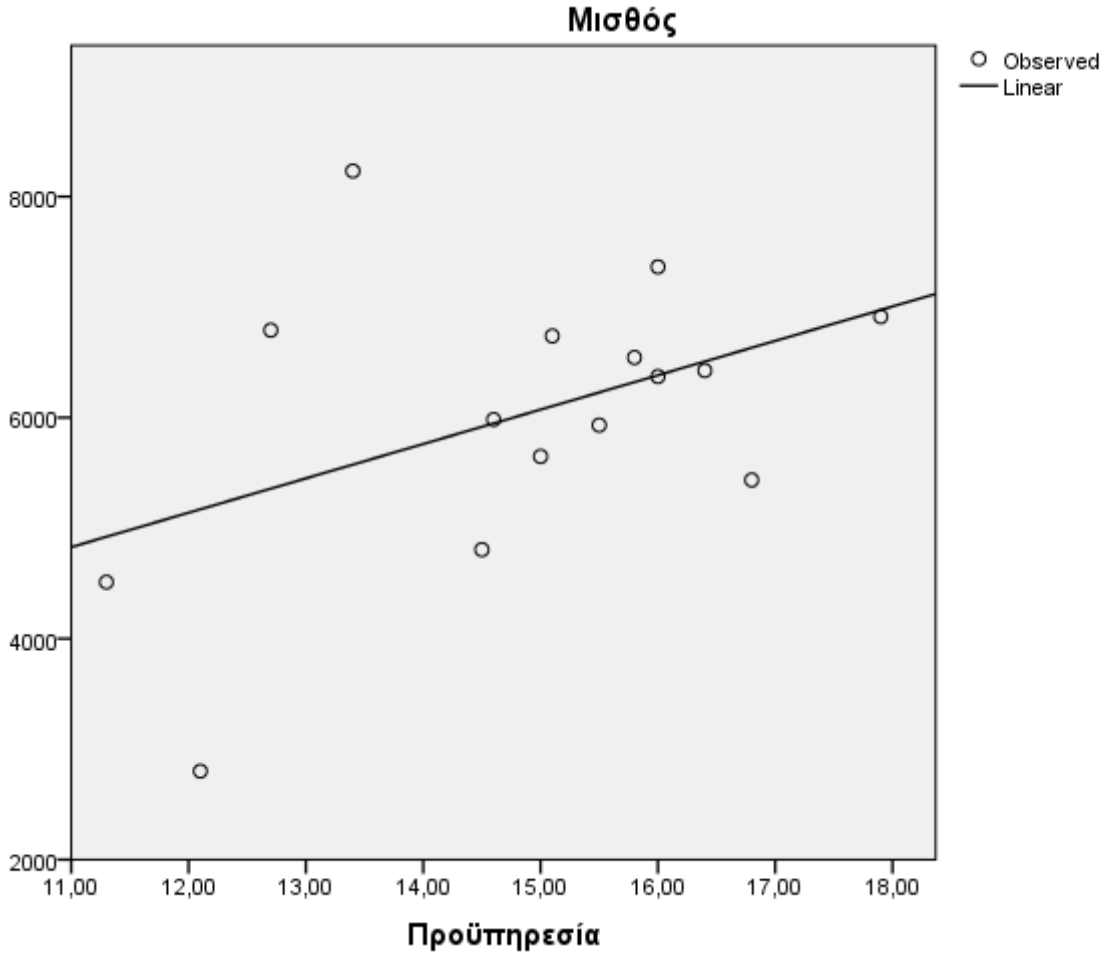
$$\hat{Y} = 1407,372 + 311,014 * X$$

Model Summary and Parameter Estimates

Dependent Variable: Μισθός

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	,190	3,043	1	13	,105	1407,372	311,014

The independent variable is Προϋπηρεσία.



Για τον ρόλο “Software Engineer” σχηματίζεται ένα δείγμα από 15 άτομα, και μέσω του SPSS προκύπτει η εξής γραμμική εξίσωση

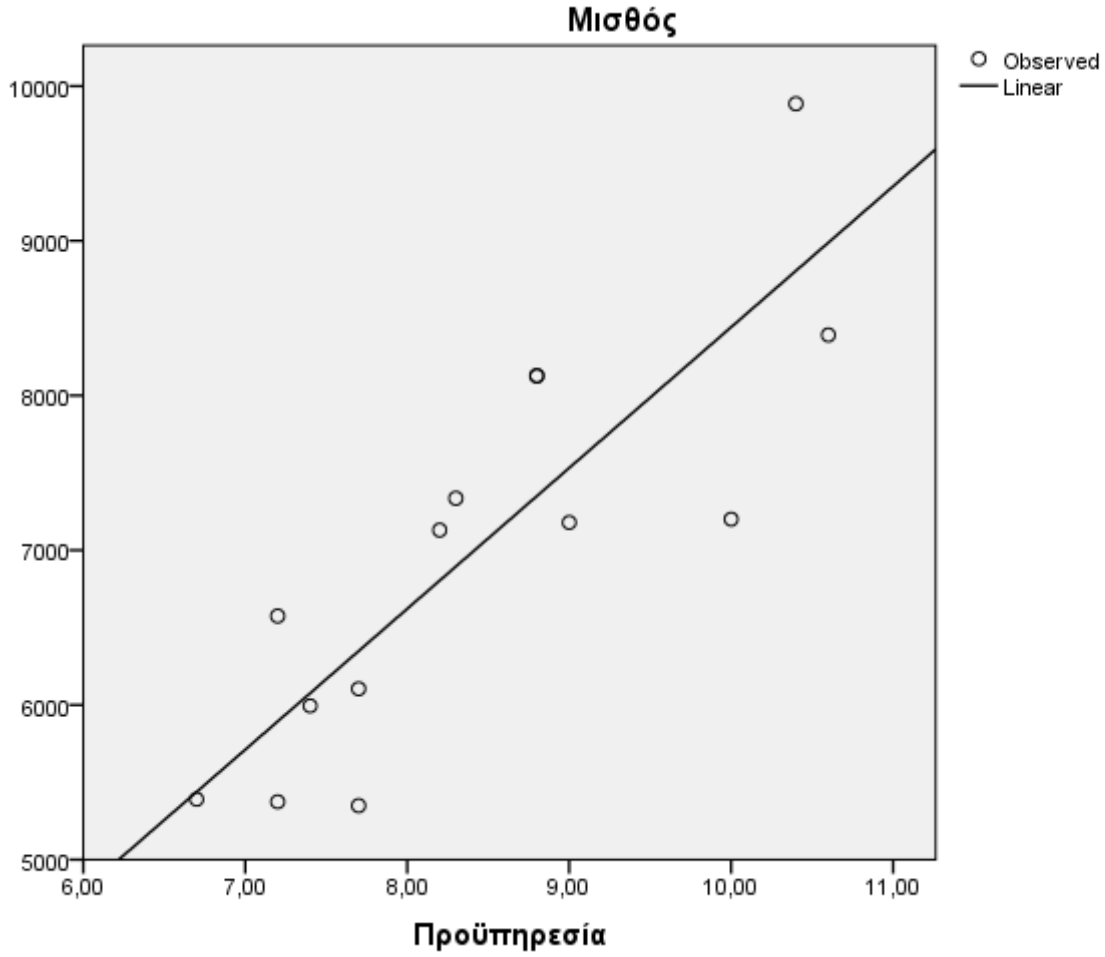
$$\widehat{Y} = -660,882 + 910,301 * X$$

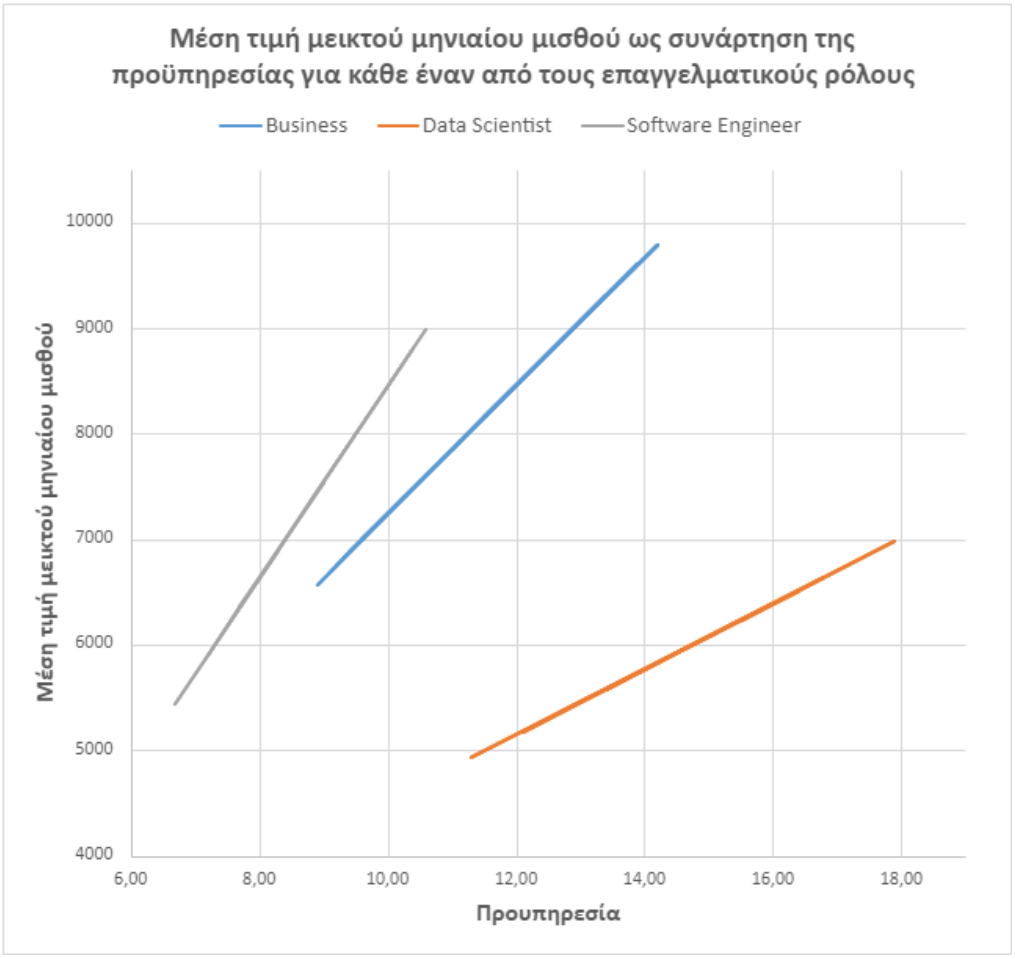
Model Summary and Parameter Estimates

Dependent Variable: Μισθός

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	,717	30,401	1	12	,000	-660,882	910,301

The independent variable is Προϋπηρεσία.


Μισθός
Observed
Linear
10000
9000
8000
7000
6000
5000
6,00
7,00
8,00
9,00
10,00
11,00
Προϋπηρεσία


Μέση τιμή μεικτού μηνιαίου μισθού ως συνάρτηση της προϋπηρεσίας για κάθε έναν από τους επαγγελματικούς ρόλους
Business
Data Scientist
Software Engineer
Μέση τιμή μεικτού μηνιαίου μισθού
10000
9000
8000
7000
6000
5000
4000
6,00
8,00
10,00
12,00
14,00
16,00
18,00
Προυπηρεσία

Θέμα 8.

Στο Worksheet Lifetime παρουσιάζονται οι χρόνοι ζωής για λάμπες LED κατά τον ποιοτικό έλεγχο μίας παρτίδας παραγωγής. Σε επίπεδο σημαντικότητας a = 0.05, ελέγξτε αν ο χρόνος ζωής ακολουθεί την εκθετική κατανομή.

H_0: οι χρόνοι ζωής για λάμπες LED ακολουθούν την εκθετική κατανομή.

H_1: οι χρόνοι ζωής για λάμπες LED δεν ακολουθούν την εκθετική κατανομή.

Για να ελέγξω αν ο χρόνος ζωής ακολουθεί την εκθετική κατανομή αρχικά

Ομαδοποιώ τις τιμές γιατί η εκθετική κατανομή είναι συνεχής.

Θέλω αυτές τις 300 τιμές θα τις ομαδοποιήσω σε 5 bins.

$$\frac{Max(X)}{10} = \frac{449.8}{10} \approx 45$$

Ο χρόνος ζωής μια LED λάμπας δεν μπορεί να είναι αρνητική, άρα ξεκινάμε από 0.

[0,45]
(45,90]
(90,135]
(135,180]
(180,225]
(225,270]
(270,315]
(315,360]
(360,405]
(405, +∞)

Function εκθετικής κατανομής

$P\{X \leq a\} = (1 - e^{-\lambda * a})$

Excel function εκθετικής κατανομής: **EXPON.DIST**

Ισχύει $\mathrm{E}[\mathrm{X}] = \frac{1}{\lambda}$ (=) $\lambda = \frac{1}{E[X]}$ όπου $\bar{X} = E[X]$

Στο excel το $\bar{X}$ βρίσκω με **AVERAGE** function

=AVERAGE(B2:B301)

AVERAGE
207,1415173

Και $\lambda = \frac{1}{207.1415173} = 0.004827617434$

Για να βρούμε την πιθανότητα για κάθε bin

Αρχικά πρέπει

[0, 45]:

$$P\{X \leq 45\} - P\{X \leq 0\} = \left(1 - e^{-\lambda * 45}\right) - \left(1 - e^{-\lambda * 0}\right) = 0.195$$

=EXPON.DIST(45;H2;TRUE)-EXPON.DIST(0;H2;TRUE)

[90,45]

$$P\{X \leq 90\} - P\{X \leq 45\} = \left(1 - e^{-\lambda * 90}\right) - \left(1 - e^{-\lambda * 45}\right) = 0.157$$

=EXPON.DIST(90;H2;TRUE)-EXPON.DIST(45;H2;TRUE)

Συνεχίζουμε την ίδια διαδικασία για τις υπόλοιπα bins. Ώσπου:

$[405, +\infty]$

$$P\{X \geq 405\} = \left(e^{-\lambda * 405}\right) = 0.141$$

=1- EXPON.DIST(405;H2;TRUE)

Αν προσθέσω τις πιθανότητες αθροίζουν στο 1. (Excel **SUM** function)

Bins	Probability
[0,45]	0,1952654315
(45,90]	0,1571368428
(90,135]	0,1264534494
(135,180]	0,101761462
(180,225]	0,08189096621
(225,270]	0,06590049136
(270,315]	0,05303240348
(315,360]	0,04267700833
(360,405]	0,03434366388
(405, +∞)	0,1415382811
SUM	1

Expected Values

$n * P\{X \leq 45\} - P\{X \leq 0\}$ =300*0.1952=58.57

$n * P\{X \leq 90\} - P\{X \leq 45\}$ =300*0.1571=47.14

Συνεχίζουμε την ίδια διαδικασία για τις υπόλοιπα bins. Ώσπου:

$n * P\{X \geq 405\}$ =300*0.141=42.46

Bins	Probability	Excpeted
[0,45]	0,1952654315	58,57962945
(45,90]	0,1571368428	47,14105283
(90,135]	0,1264534494	37,93603481
(135,180]	0,101761462	30,5284386
(180,225]	0,08189096621	24,56728986
(225,270]	0,06590049136	19,77014741
(270,315]	0,05303240348	15,90972104
(315,360]	0,04267700833	12,8031025
(360,405]	0,03434366388	10,30309916
(405, +oo)	0,1415382811	42,46148433
SUM	1	300

Το άθροισμα των expected values αθροίζουν στο 300 (μέγεθος δείγματος)

Observed Values

Στο excel για τα observed values χρησιμοποίησα την **COUNTIF** και **COUNTIFS**

[0, 45]

=COUNTIF(B2:B301; "<45")

[45, 90]

=COUNTIFS(B2:B301; "<90"; B2:B301; ">45")

Συνεχίζουμε την ίδια διαδικασία για τις υπόλοιπα bins. Ώσπου:

$[405, +\infty]$

=COUNTIF(B2:B301; ">405")

Bins	Probability	Excpeted	Observed
[0,45]	0,1952654315	58,57962945	42
(45,90]	0,1571368428	47,14105283	28
(90,135]	0,1264534494	37,93603481	29
(135,180]	0,101761462	30,5284386	34
(180,225]	0,08189096621	24,56728986	35
(225,270]	0,06590049136	19,77014741	26
(270,315]	0,05303240348	15,90972104	36
(315,360]	0,04267700833	12,8031025	23
(360,405]	0,03434366388	10,30309916	23
(405, +∞)	0,1415382811	42,46148433	24
SUM	1	300	300

Το άθροισμα των observed values αθροίζουν στο 300 (μέγεθος δείγματος)

$H_{0:}$ Η πιθανότητα εμφάνισης κάθε ενός από τα πέντε διαστήματα στον πληθυσμό είναι:

$$P_1 = 0.195, P_2 = 0.157, P_3 = 0.126, P_4 = 0.101, P_5 = 0.081$$

$$P_6 = 0.065, P_7 = 0.053, P_8 = 0.042, P_9 = 0.034, P_{10} = 0.141$$

H_1: Τουλάχιστον δύο από τις πιθανότητες εμφάνισης των διαστημάτων στον πληθυσμό διαφέρουν από αυτές της H_0.

Υπολογισμός Κριτηρίου Ελέγχου

$$X^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = \frac{(42 - 58.57)^2}{58.57} + \frac{(28 - 47.14)^2}{47.14} + \cdots$$

$$+ \frac{(23-10.3)^2}{10.3} + \frac{(24-42.46)^2}{42.46} = 4.69 + 7.77 + \ldots + 15.64 + 8.02$$

$$= 78.521$$

Bins	(Observed-Expected)^2/Expected
[0,45]	4,692486371
(45,90]	7,771992382
(90,135]	2,104930536
(135,180]	0,3947708792
(180,225]	4,430339748
(225,270]	1,963114514
(270,315]	25,3693517
(315,360]	8,121212704
(360,405]	15,64687365
(405, +oo)	8,026719014
SUM	78,5217915

$X^2{}_{0.95,8} = 15{,}5$

$\nu = k - m - 1 = 10 - 1 - 1 = 8$

$m = 1$, προσέγγιση της $\bar{X}$ με **AVERAGE**

$X^2{}_{0.95,8} < X^2$, το κριτήριο ελέγχου βρίσκεται στη ζώνη απόρριψης της H_0.

Επομένως, σε επίπεδο σημαντικότητας 0.05, τα δειγματικά δεδομένα δεν είναι συμβατά με την εκθετική κατανομή.