ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΠΛΗΡΟΦΟΡΙΚΗ»

Πιθανότητες και Στατιστική
Εργασία Εξαμήνου

Αναγνωστόπουλος Βασίλης - Θάνος

ΑΘΗΝΑ, 2014

Περιεχόμενα

1 Εισαγωγή - Αντικείμενο της άσκησης 1

2 Μέρος Α 1

3 Μέρος Β 8

4 Μέρος Γ 13

5 Μέρος Δ 16

6 Μέρος Ε 20

Βιβλιογραφία 21

1. Εισαγωγή - Αντικείμενο της άσκησης

Τα δεδομένα στο αρχείο views.sav αποτελούν τυχαίο δείγμα 52 σελίδων και περιέχουν τις ακόλουθες μεταβλητές:

Όνομα μεταβλητής	**Περιγραφή μεταβλητής**
Country	Χώρα προέλευσης (1 = ελληνική, 0 = όχι ελληνική)
Subject	Θεματολογία της ιστοσελίδας (1 = Αθλητικά, 2 = Πολιτικά, 3 = Lifestyle)
News	Ημερήσιος αριθμός νέων αναρτήσεων
Yr	Παλαιότητα της ιστοσελίδας (1 = λειτουργεί λιγότερο από 2 έτη, 0 = διαφορετικά)
Journalists	Αριθμός δημοσιογράφων που απασχολούνται στη συγκεκριμένη ιστοσελίδα
Views	Ετήσιος αριθμός επισκέψεων (views) σε συγκεκριμένη ιστοσελίδα

Πίνακας 1.1: Οι μεταβλητές του αρχείου views.sav.

Από το αρχικό αρχείο αφαιρέθηκε η 2η παρατήρηση με τιμές (0,1,12,1,22,35350)

2. Μέρος Α

- Να υπολογισθεί ή μέση τιμή και η τυπική απόκλιση των ετήσιων αριθμών επισκέψεων των ιστοσελίδων. Να δοθεί η ερμηνεία των αποτελεσμάτων.

Μέσος όρος ή αλλιώς δειγματική μέση τιμή ενός συνόλου ν παρατηρήσεων είναι ένα μέτρο θέσης, δηλαδή δείχνει σχετικά τις θέσεις των αριθμών στους οποίους αναφέρεται. Γενικά, ορίζεται ως το άθροισμα των παρατηρήσεων δια του πλήθους αυτών. Είναι δηλαδή η μαθηματική πράξη ανεύρεσης της «μέσης απόστασης» ανάμεσα σε δύο ή περισσότερους αριθμούς. Η μέση τιμή συμβολίζεται με $\bar{x}$. Γενικός τύπος της μέσης τιμής είναι [5]:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{1}{n}(t_1 + \cdots + t_n) \tag{1}$$

όπου t_i η i παρατήρηση και n το πλήθος των παρατηρήσεων

Η διακύμανση ή διασπορά μίας τυχαίας μεταβλητής x συμβολίζεται συνήθως με $Var[x]$ και δηλώνει πόσο συγκεντρωμένες γύρω από τη μέση τιμή είναι οι τιμές της τυχαίας μεταβλητής. Η θετική τετραγωνική ρίζα της διακύμανσης ονομάζεται τυπική απόκλιση

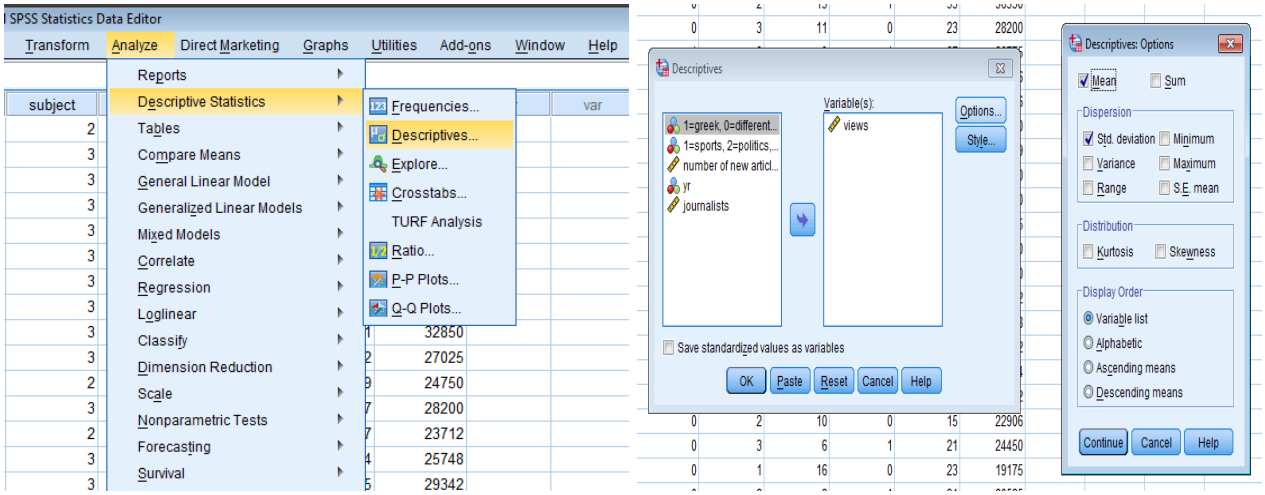
και συμβολίζεται με σ. Ο γενικός τύπος της απόκλισης είναι [4]:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{x})^2} \tag{2}$$

Για τον υπολογισμό τους στο SPSS πηγαίνουμε στο μενού: Analyze|Descriptive Statistics|Descriptives (βλ. σχήμα 2.1) και προκύπτει ο πίνακας 2.1.

	N	Mean	Std. Deviation
views	51	23571.14	5743.964
Valid N (listwise)	51		

Πίνακας 2.1: Η μέση τιμή και η τυπική απόκλιση των ετήσιων αριθμών επισκέψεων των ιστοσελίδων.



Σχήμα 2.1: Το μενού Analyze|Descriptive Statistics|Descriptives του SPSS

Παρατηρούμε ότι η μέση τιμή είναι ίση με $23571,14$. Αυτό πρακτικά σημαίνει ότι η κεντρική τάση των επισκέψεων των ιστοσελίδων είναι $23571,14$. Πρόσθετα η τυπική απόκλιση του δείγματος των 51 δειγμάτων ισούται με 5743.964. Η τυπική απόκλιση εκφράζει το βαθμό διασποράς των επισκέψεων των ιστοσελίδων, δηλαδή περιγράφει το αν το δείγμα των βαθμολογίων αποτελείται από παρατηρήσεις που έχουν κοντινές ή μακρινές αποστάσεις μεταξύ τους [7].

- Να υπολογισθεί η διάμεσος, τα τεταρτημόρια, το 40% ποσοστημόριο και η κορυφή των ετησίων αριθμών επισκέψεων των ιστοσελίδων. Να δοθεί η ερμηνεία των αποτελεσμάτων.

Διάμεσος ενός συνόλου ν παρατηρήσεων είναι η αριθμητική τιμή που διαχωρίζει το υψηλότερο ήμισυ ενός δείγματος δεδομένων, έναν πληθυσμό ή μία κατανομή πιθανοτήτων από το κάτω μισό, όταν αυτές έχουν διαταχθεί σε αύξουσα σειρά [1].

Τεταρτημόριο είναι ένα μέτρο που χρησιμοποιείται στην στατιστική και υποδηλώνει την τιμή κάτω από την οποία ένα δεδομένο ποσοστό παρατηρήσεων βρίσκονται κάτω από

αυτή την τιμή [3].

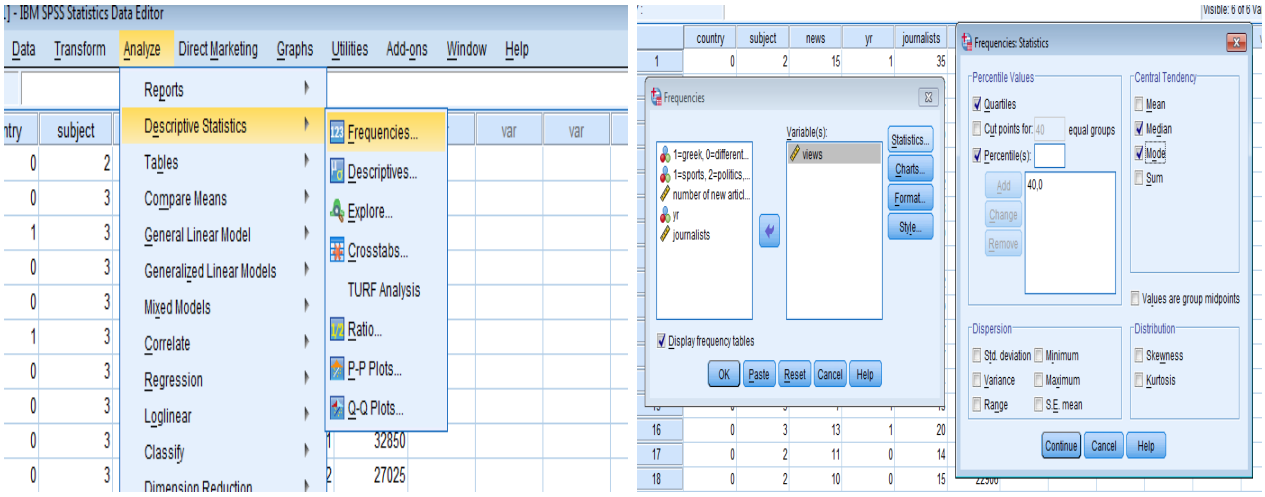
Κορυφή είναι η τιμή που εμφανίζεται πιο συχνά σε ένα σύνολο δεδομένων [2].

Για τον υπολογισμό τους στο SPSS πηγαίνουμε στο μενού: Analyze|Descriptive Statistics|Frequencies (βλ. σχήμα 2.2) και προκύπτει ο πίνακας 2.2.

N	Valid	51
	Missing	0
Median		23713.00
Mode		28200
Percentiles	25	18075.00
Percentiles	40	21479.80
Percentiles	50	23713.00
Percentiles	75	27025.00

Πίνακας 2.2: Η διάμεσος, τα τεταρτημόρια, το ποσοστημόριο και η κορυφή των ετήσιων αριθμών επισκέψεων των ιστοσελίδων.

Παρατηρούμε ότι η διάμεσος (median) είναι ίση με 23713, που σημαίνει ότι οι 25 ιστοσελίδες έχουν μέχρι 23713 επισκέψεις ενώ οι υπόλοιπες παραπάνω. Η κορυφή των παρατηρήσεων είναι 28200, που σημαίνει ότι είναι η πιο συχνά εμφανιζόμενη τιμή. Τέλος το πρώτο τεταρτημόριο είναι ίσο με 18075 (που σημαίνει ότι το 1/4 των ιστοσελίδων έχουν επισκέψεις μέχρι 18075) και ομοίως και για τα υπόλοιπα. Το ποσοστημόριο 40% ισούται με 21479.80 που ομοίως σημαίνει ότι το 30% των σελίδων έχουν επισκέψεις μέχρι και 21479.80 .



Σχήμα 2.2: Το μενού Analyze|Descriptive Statistics|Frequencies του SPSS

- Να ορισθεί κατάλληλα μια νέα μεταβλητή (Views_2), η οποία να ομαδοποιεί τις ιστοσελίδες σε 4 κατηγόριες ανάλογα με τον ετήσιο αριθμό επισκέψεων τους ως εξής:

  1η ομάδα: ιστοσελίδες με ετήσιο αριθμό επισκέψεων μέχρι 22000

  2η ομάδα: ιστοσελίδες με ετήσιο αριθμό επισκέψεων πάνω από 22000 και μέχρι 28000

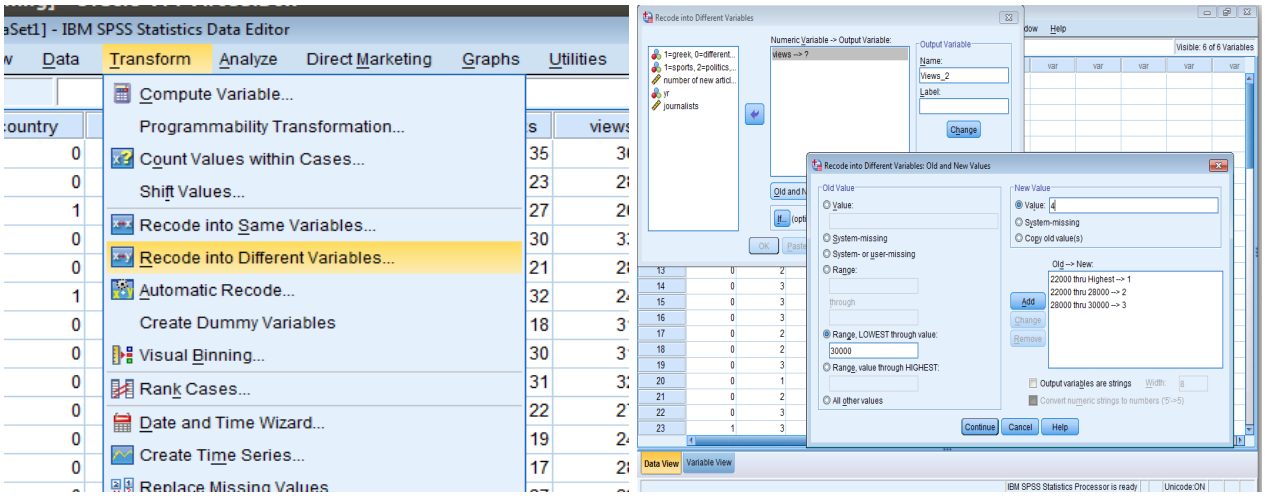
  3η ομάδα: ιστοσελίδες με ετήσιο αριθμό επισκέψεων πάνω από 28000 και μέχρι 30000

  4η ομάδα: ιστοσελίδες με ετήσιο αριθμό επισκέψεων πάνω από 30000

  – Να κατασκευασθεί ο πίνακας συχνοτήτων και το κυκλικό διάγραμμα βάσει της νέας μεταβλητής

  – Τί ποσοστό των ιστοσελίδων ανήκουν στην 3η ομάδα;

Για τον ορισμό της νέας μεταβλητής στο SPSS πηγαίνουμε στο μενού: Transform | Recode Into different Variables... (βλ. σχήμα 2.3).



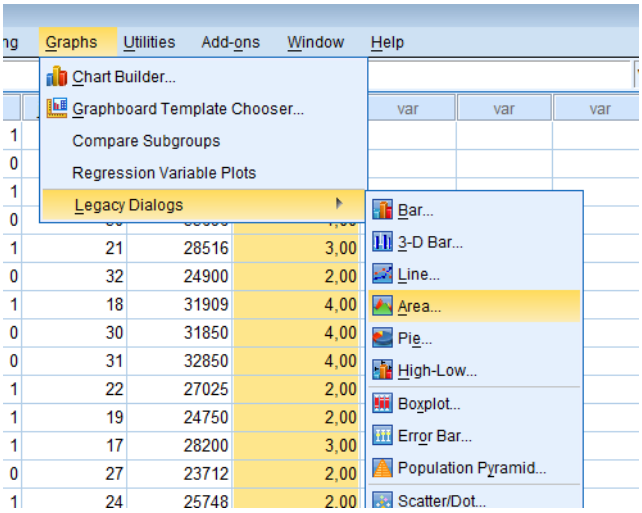
Σχήμα 2.3: Το μενού Transform | Recode Into different Variables... του SPSS

Μετά ομοίως με το προηγούμενο ερώτημα υπολογίζουμε τον πίνακα συχνοτήτων. Για τον υπολογισμό τους στο SPSS πηγαίνουμε στο μενού: Analyze|Descriptive Statistics|Frequencies και προκύπτει ο πίνακας 2.3.

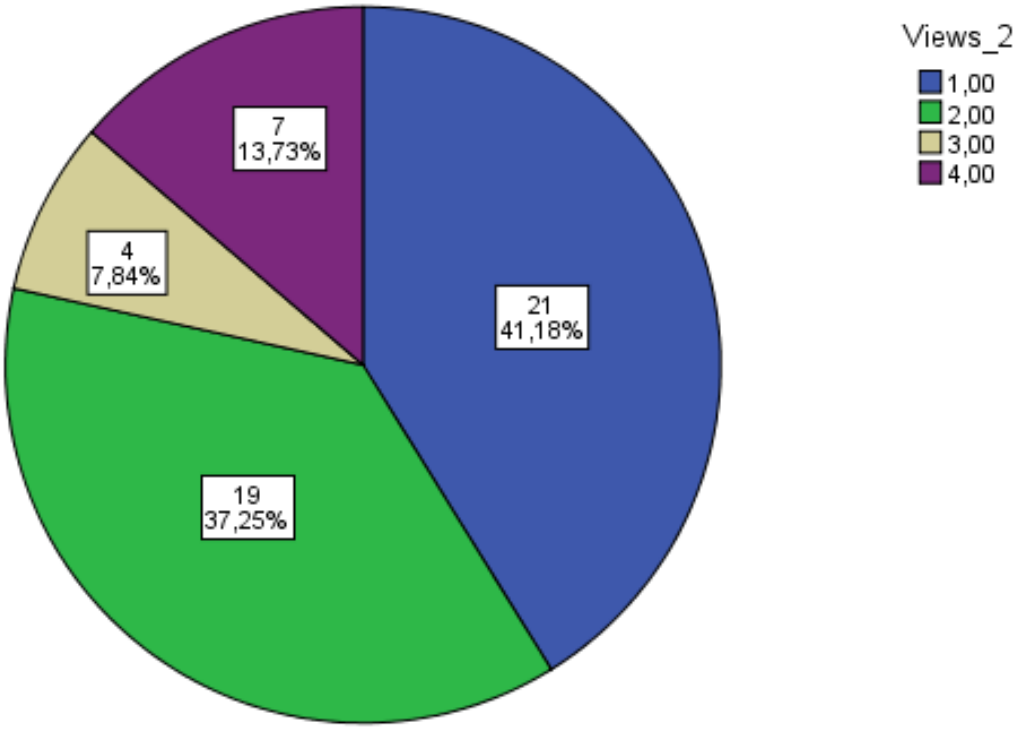
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	21	41.2	41.2	41.2
	2.00	19	37.3	37.3	78.4
	3.00	4	7.8	7.8	86.3
	4.00	7	13.7	13.7	100.0
	Total	51	100.0	100.0	

Πίνακας 2.3: Ο πίνακας συχνοτήτων της μεταβλητής Views_2.

Επομένως για την δημιουργία του κυκλικού διαγράμματος πηγαίνουμε στο μενού: Graphs|Legacy Dialogs|Pie (βλ. σχήμα 2.4) και προκύπτει το σχήμα 2.5.



Σχήμα 2.4: Το μενού Graphs|Legacy Dialogs|Pie του SPSS.



Σχήμα 2.5: Το κυκλικό διάγραμμα που προκύπτει από την μεταβλητή Views_2.

Επομένως το ποσοστό που ανήκει στην 3η ομάδα είναι 7.84% .

- Να συγκριθούν ως προς τη μεταβλητότητα που παρουσιάζουν οι 4 ομάδες ιστοσελίδων που έχουν δημιουργηθεί βάσει της νέας μεταβλητής (Views_2). Σχολιάστε τα αποτελέσματα.

Τα μέτρα διασποράς δίνουν πληροφορίες για την μεταβλητότητα, δηλαδή το "άπλωμα" των παρατηρήσεων σε ένα σύνολο δεδομένων. Με αυτά δίνεται η δυνατότητα να εξαχθούν άμεσα συμπεράσματα σχετικά με την συμπεριφορά των υποκείμενων τιμών. [6]

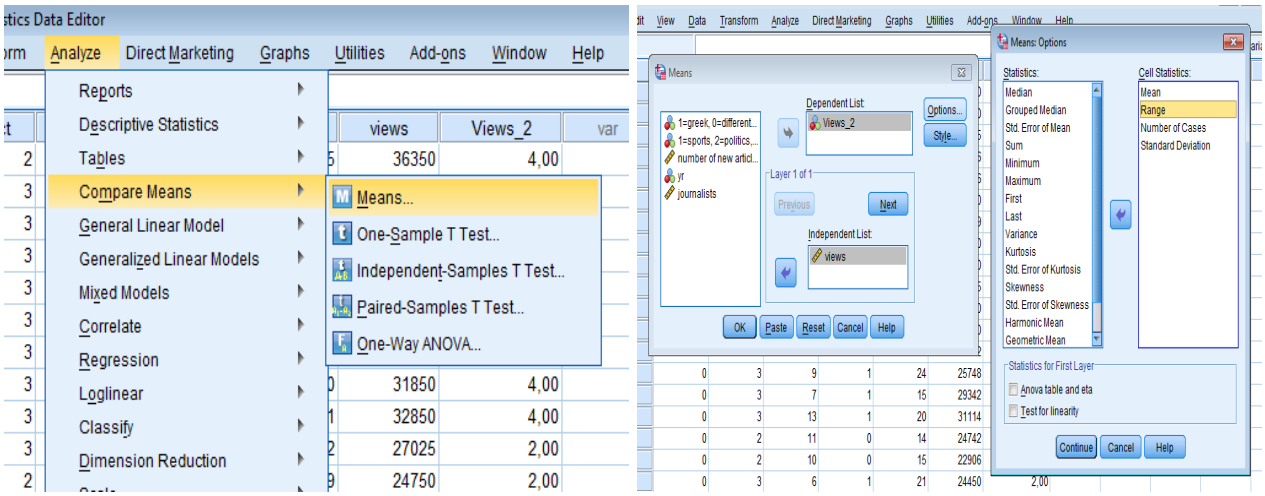
Μερικά από τα βασικότερα μέτρα διασποράς ή μεταβλητότητας είναι:

Εύρος (αγγλ. Range): Το εύρος R ενός δείγματος ορίζεται ως η διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής αυτού. Δηλαδή:

$$R = t_max - t_min \tag{3}$$

Τυπική απόκλιση: Έχει ορισθεί πιο πάνω.

Για τον υπολογισμό των μέσων τιμών και του εύρους για τις 4 ομάδες στο SPSS πηγαίνουμε στο μενού: Analyze|Compare Means|Means (βλ. σχήμα 2.6) και προκύπτει ο πίνακας 2.4.



Σχήμα 2.6: Το μενού Analyze|Compare Means|Means του SPSS

Views_2	Mean	Range	N	Std. Deviation
1,00	18056.52	6600	21	2046.452
2,00	24887.84	5509	19	1440.729
3,00	28564.50	1142	4	539.314
4,00	33687.71	3931	7	2580.070
Total	23571.14	23045	51	5743.964

Πίνακας 2.4: Οι μεταβλητότητες που παρουσιάζουν οι 4 ομάδες ιστοσελίδων.

Όπως παρατηρούμε η ομάδες δεν είναι τόσο ομοιογενές μιας και το εύρος τιμών τους είναι αρκετά διαφορετικό και ο αριθμός των παρατηρήσεων που περιλαμβάνονται στις

4 ομάδες είναι διαφορετικός. Τέλος οι τυπικές τους αποκλίσεις διαφέρουν αρκετές μεταξύ τους, επομένως ούτε μέσα στην κάθε ομάδα οι τιμές είναι ομοιογενές και αυτές απλώνονται αρκετά μέσα στην κάθε ομάδα.

3. Μέρος Β

- Να εξετασθεί σε επίπεδο σημαντικότητας 5% αν οι ελληνικές ιστοσελίδες έχουν την ίδια μέση ετήσια επισκεψιμότητα με τις όχι ελληνικές ιστοσελίδες.

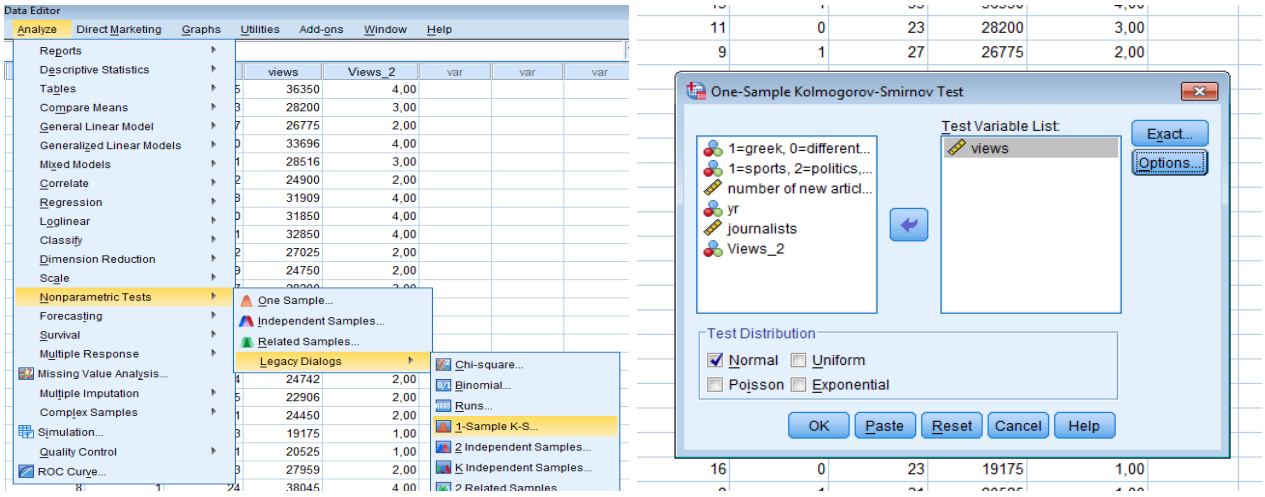
Οι δύο υποθέσεις που έρχονται σε αντίθεση σύμφωνα με την εκφώνηση είναι οι ακόλουθες:

$$H_0 : \mu_g = \mu_f, H_1 : \mu_g \neq \mu_f \tag{4}$$

όπου μ_g, μ_f είναι οι η επισκεψιμότητα των ελληνικών και ξένων ιστοσελίδων αντίστοιχα.

Προκειμένου να εφαρμόσουμε παραμετρικό έλεγχο για την σύγκριση της επισκεψιμότητας των ιστοσελίδων θα πρέπει πρώτα να εξετάσουμε αν τα δεδομένα που διαθέτουμε προσαρμόζονται ικανοποιητικά στην κανονική κατανομή.

Για τον έλεγχο στο SPSS πηγαίνουμε στο μενού: Analyze|NonParametric test|1 Sample K-S (βλ. σχήμα 3.1) και λαμβάνουμε τον πίνακα 3.1.



Σχήμα 3.1: Το μενού Analyze|NonParametric test|1 Sample K-S του SPSS

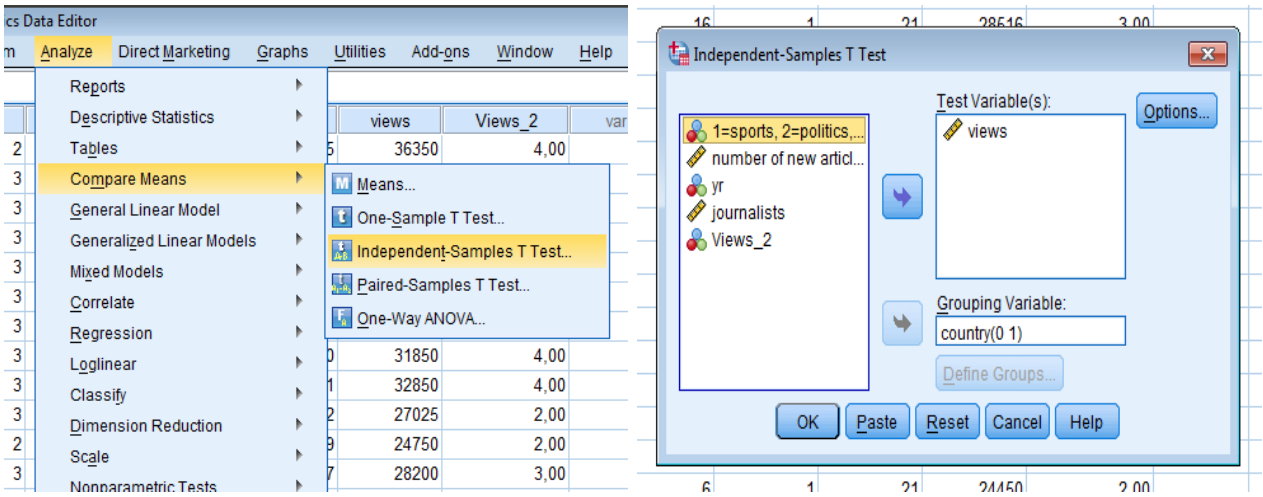
		views
N		51
Normal Parameters	Mean	23571,14
	Std. Deviation	5743,964
Most Extreme Differences	Absolute	,095
	Positive	,095
	Negative	-,068
Test Statistic		,095
Asymp. Sig. (2-tailed)		,200

Πίνακας 3.1: Οι μεταβλητότητες που παρουσιάζουν οι 4 ομάδες ιστοσελίδων.

Η τιμή p-value (Asymp. Sigm.(2-tailed)) για τον έλεγχο της κανονικότητας των δεδομέ-

νων είναι ίση με $0.200 > 0.05$. Συνεπώς αποδεχόμαστε τη μηδενική υπόθεση της καλής προσαρμογής των δεδομένων στην κανονική κατανομή.

Στη συνέχεια, για τον έλεγχο της υπόθεσης στο SPSS πηγαίνουμε στο μενού: Analyze | Compare means | Independent samples T-test (βλ. σχήμα 3.2) και προκύπτει ο πίνακας 3.2.



Σχήμα 3.2: Το μενού Analyze | Compare means | Independent samples T-test του SPSS

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
views	Equal variances assumed	,016	,901	1,726	49	,091	3051,722	1767,651	-500,505	6603,949
	Equal variances not assumed			1,631	21,144	,118	3051,722	1871,191	-838,022	6941,466

Πίνακας 3.2: Ο πίνακας που προκύπτει από το μενού Analyze | Compare means | Independent samples T-test του SPSS

Από τον πίνακα 3.2 παρατηρούμε ότι $p - value = 0.901 > 0.05$, συνεπώς (σε επίπεδο σημαντικότητας 5%) δεχόμαστε την μηδενική υπόθεση, που σημαίνει ότι οι ελληνικές ιστοσελίδες έχουν την ίδια μέση ετήσια επισκεψιμότητα με τις ξένες ιστοσελίδες.

- Να εξετασθεί σε επίπεδο σημαντικότητας 5% αν οι ιστοσελίδες που λειτουργούν τουλάχιστον 2 έτη, έχουν τον ίδιο μέσο ετήσιο αριθμό επισκέψεων με ιστοσελίδες που λειτουργούν λιγότερο από 2 έτη.

Οι δύο υποθέσεις που έρχονται σε αντίθεση σύμφωνα με την εκφώνηση είναι οι ακόλουθες:

$$H_0 : \mu_{<2} = \mu_{>2}, H_1 : \mu_{<2} \neq \mu_{>2} \quad (5)$$

όπου $\mu_{<2}, \mu_{>2}$ είναι οι η επισκεψιμότητα των σελίδων που λειτουργούν τουλάχιστον 2 έτη και οι σελίδες που λειτουργούν λιγότερο από 2 έτη αντίστοιχα. Από το προηγούμενο

ερωτήμα γνωρίζουμε ότι τα δεδομένα ακολουθούν κανονική κατανομή επομένως δεν χρειάζεται να κάνουμε πάλι έλεγχο.

Στη συνέχεια, για τον έλεγχο της υπόθεσης στο SPSS, ομοίως με πριν, πηγαίνουμε στο μενού: Analyze | Compare means | Independent samples T-test. Το μόνου που αλλάζει είναι το Grouping Variable σε yr(0 1) και προκύπτει ο πίνακας 3.3.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
views	Equal variances assumed	4,580	,037	1,087	49	,282	1851,324	1703,121	-1571,227	5273,874
	Equal variances not assumed			1,209	42,349	,234	1851,324	1531,822	-1239,265	4941,912

Πίνακας 3.3: Ο πίνακας που προκύπτει από το μενού Analyze | Compare means | Independent samples T-test του SPSS

Από τον πίνακα 3.3 παρατηρούμε ότι $p - value = 0.037 < 0.05$, συνεπώς (σε επίπεδο σημαντικότητας 5%) απορρίπτουμε τη μηδενική υπόθεση, που σημαίνει ότι οι ιστοσελίδες που λειτουργούν τουλάχιστον 2 έτη, δεν έχουν τον ίδιο μέση ετήσιο αριθμό επισκέψεων με τις ιστοσελίδες που λειτουργούν λιγότερο από 2 έτη.

- Να εξετασθεί σε επίπεδο σημαντικότητας 5% αν ο μέσος ετήσιος αριθμός επισκέψεων των ιστοσελίδων είναι στατιστικά ίσος με 25000 ή όχι.

Οι δύο υποθέσεις που έρχονται σε αντίθεση σύμφωνα με την εκφώνηση είναι οι ακόλουθες:

$$H_0 : \mu_A = \mu_0, H_1 : \mu_A \neq \mu_0 \tag{6}$$

όπου $\mu_0 = 25000$ και μ_A είναι η άγνωστη επισκεψιμότητα των ιστοσελίδων αντίστοιχα. Από τα προηγούμενα ερωτήματα γνωρίζουμε ότι τα δεδομένα ακολουθούν κανονική κατανομή επομένως δεν χρειάζεται να κάνουμε πάλι έλεγχο.

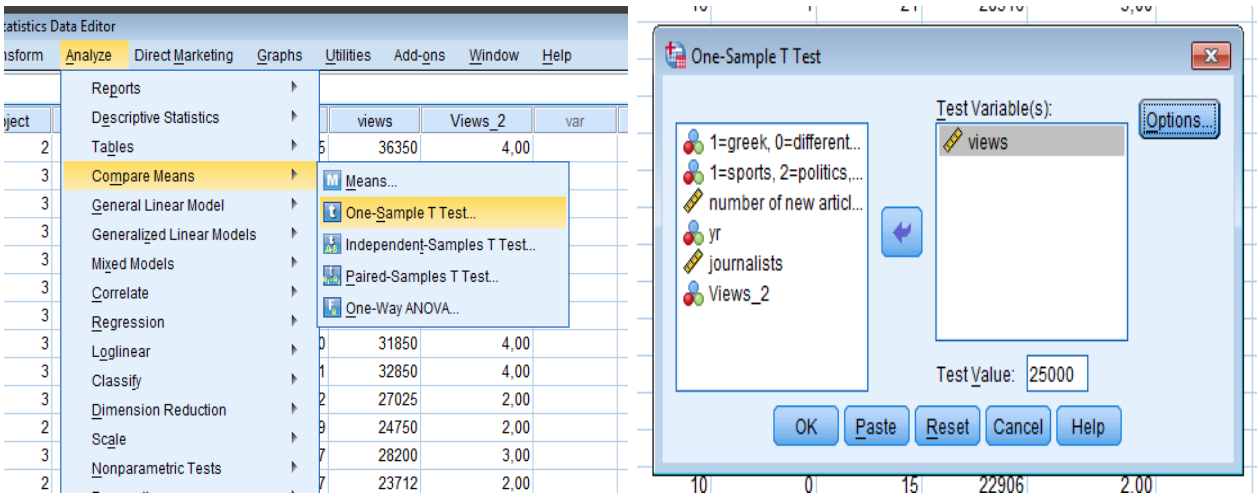
Στη συνέχεια για τον έλεγχο της υπόθεσης στο SPSS πηγαίνουμε στο μενού: Analyze | Compare means | one sample T-Test (βλ. σχήμα 3.3) και προκύπτει ο πίνακας 3.4.

One-Sample Test

	Test Value = 25000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
views	-1,776	50	,082	-1428,863	-3044,38	186,65

Πίνακας 3.4: Ο πίνακας που προκύπτει από το μενού Analyze | Compare means | one sample T-Test του SPSS

Από τον πίνακα 3.4 παρατηρούμε ότι $p - value = 0.082 > 0.05$. Συνεπώς (σε επίπεδο



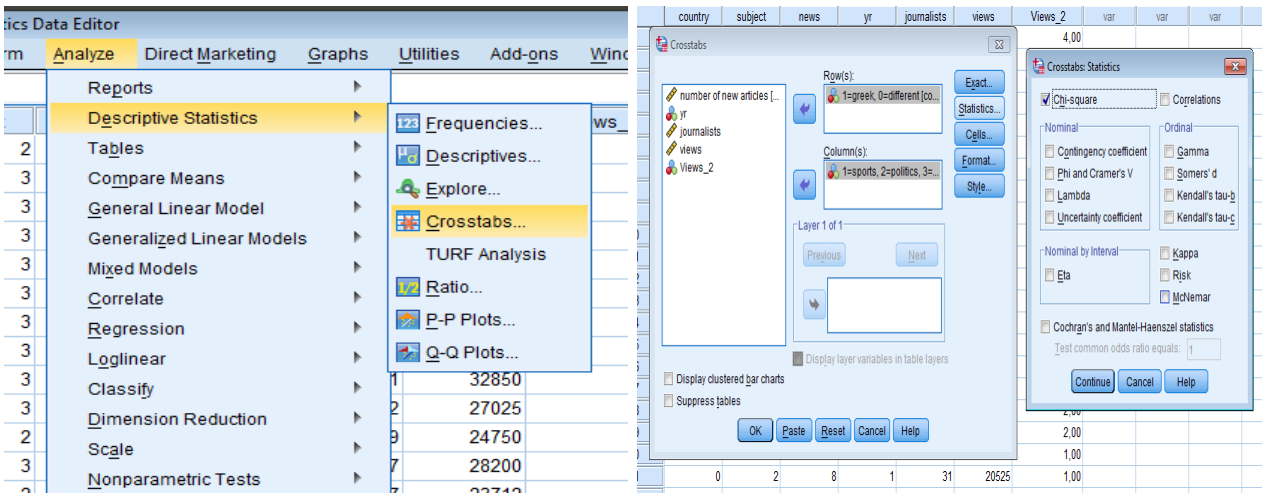
Σχήμα 3.3: Το μενού Analyze | Compare means | Independent samples T-test του SPSS

σημαντικότητας 5%) δεν απορρίπτουμε τη μηδενική υπόθεση, γεγονός που σημαίνει ότι ο μέσος ετήσιος αριθμός επισκέψεων των ιστοσελίδων είναι στατιστικά ίσος με 25000.

- Να εξετασθεί σε επίπεδο σημαντικότητας 5% αν ο παράγοντας Country επηρεάζει τη θεματολογία της ιστοσελίδας.

Οι δύο υποθέσεις που έρχονται σε αντιπαράθεση σύμφωνα με την εκφώνηση είναι οι: H_0 και H_1 όπου η H_0 : η θεματολογία της ιστοσελίδας είναι ανεξάρτητη από τον παράγοντα Country και H_1 η θεματολογία της ιστοσελίδας εξαρτάται από τον παράγοντα Country.

Για τον έλεγχο της υπόθεσης στο SPSS πηγαίνουμε στο μενού: Analyze | Descriptive Statistics | Crosstabs (βλ. σχήμα 3.4) και προκύπτει ο πίνακας 3.5 και 3.6.



Σχήμα 3.4: Το μενού Analyze | Compare means | Independent samples T-test του SPSS

Από τον πίνακα 3.6 παρατηρούμε ότι $p - value = 0.113 > 0.05$, συνεπώς (σε επίπεδο σημαντικότητας 5%) δεν απορρίπτουμε την μηδενική υπόθεση, που σημαίνει ότι η θεματολογία της ιστοσελίδας είναι ανεξάρτητη από τον παράγοντα Country.

1=greek, 0=different * 1=sports, 2=politics, 3=lifestyle Crosstabulation

			1=sports, 2=politics, 3=lifestyle			
			1	2	3	Total
1=greek, 0=different	0	Count	10	13	14	37
		% within 1=greek, 0=different	27,0%	35,1%	37,8%	100,0%
		% within 1=sports, 2=politics, 3=lifestyle	55,6%	86,7%	77,8%	72,5%
		% of Total	19,6%	25,5%	27,5%	72,5%
	1	Count	8	2	4	14
		% within 1=greek, 0=different	57,1%	14,3%	28,6%	100,0%
		% within 1=sports, 2=politics, 3=lifestyle	44,4%	13,3%	22,2%	27,5%
		% of Total	15,7%	3,9%	7,8%	27,5%
Total		Count	18	15	18	51
		% within 1=greek, 0=different	35,3%	29,4%	35,3%	100,0%
		% within 1=sports, 2=politics, 3=lifestyle	100,0%	100,0%	100,0%	100,0%
		% of Total	35,3%	29,4%	35,3%	100,0%

Πίνακας 3.5: Ο πίνακας που προκύπτει από το μενού Analyze | Descriptive Statistics | Crosstabs του SPSS (1)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,358[a]	2	,113
Likelihood Ratio	4,364	2	,113
Linear-by-Linear Association	2,188	1	,139
N of Valid Cases	51		

a. 3 cells (50,0%) have expected count less than 5. The minimum expected count is 4,12.

Πίνακας 3.6: Ο πίνακας που προκύπτει από το μενού Analyze | Descriptive Statistics | Crosstabs του SPSS (2)

4. Μέρος Γ

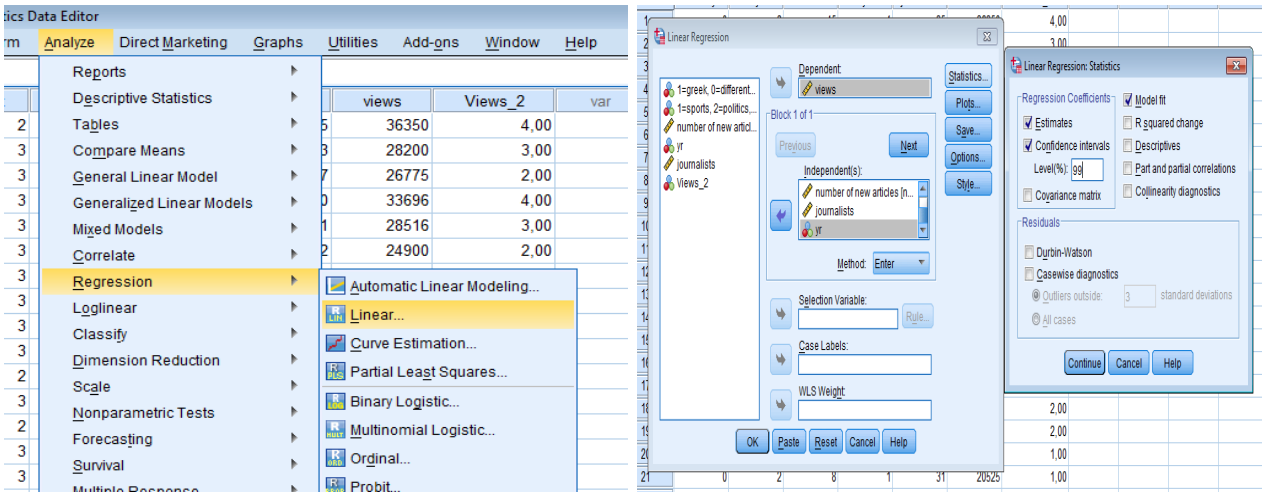
- Να εξετασθεί αν η μεταβλητή views (Y) εξαρτάται γραμμικά από τις μεταβλητές country, subject, news, journalist, yr. Να βρεθεί το βέλτιστο γραμμικό μοντέλο (σε επίπεδο σημαντικότητας 1%) και να δοθεί η γραμμική εξίσωση που αντιστοιχεί σε αυτό.

Ο έλεγχος για την ύπαρξη γραμμικής σχέσης ανάμεσα στις μεταβλητές ισοδυναμεί με τον ακόλουθο στατιστικό έλεγχο

$$H_0 : \beta = 0, H_1 : \beta \neq 0 \tag{7}$$

Οι δύο υποθέσεις που έρχονται σε αντιπαράθεση σύμφωνα με την εκφώνηση είναι οι: H_0 και H_1 όπου η H_0 : η μεταβλητή είναι γραμμικά ανεξάρτητη από τον παράγοντα και H_1 η μεταβλητή είναι γραμμικά εξαρτημένη από την μεταβλητή views.

Για τον έλεγχο της υπόθεσης στο SPSS πηγαίνουμε στο μενού: Analyze | Regression | Linear (βλ. σχήμα 4.1) και προκύπτει ο πίνακας 4.1.



Σχήμα 4.1: Το μενού Analyze | Regression | Linear του SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	99,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	11819,566	1322,158		8,940	,000	8263,510	15375,621
	1=greek, 0=different	457,290	986,781	,036	,463	,645	-2196,741	3111,321
	1=sports, 2=politics, 3=lifestyle	4421,575	681,363	,653	6,489	,000	2588,992	6254,157
	number of new articles	416,238	111,013	,365	3,749	,001	117,659	714,816
	yr	-426,137	1003,676	-,035	-,425	,673	-3125,608	2273,334
	journalists	3,461	70,877	,006	,049	,961	-187,169	194,091

a. Dependent Variable: views

Πίνακας 4.1: Ο πίνακας που προκύπτει από το μενού Analyze | Regression | Linear του SPSS.

Η απόρριψη ή αποδοχή της μηδενικής υπόθεσης θα βασιστεί στο p-value του πίνακα 4.1. Αν το $p - value < 0.01$ τότε αποδεχόμαστε την H_1, δηλαδή ότι ο παράγοντας είναι σημαντικός. Αν το $p - value > 0.01$ τότε αποδεχόμαστε την H_0. Συνεπώς (σε επίπεδο σημαντικότητας 1%) οι σημαντικοί παράγοντες είναι οι subject, news.

Επομένως κρατώντας μόνο τους σημαντικούς παράγοντες προκύπτει ο πίνακας 4.2. Οι συντελεστές της γραμμικής εξίσωσης δίνονται στην στήλη B του πίνακα. Άρα η εξίσωση είναι η $Y_{views} = 11739.834 + X_{subject} \cdot 4411.963 + X_{news} \cdot 415.654$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	99,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	11739,834	987,137		11,893	,000	9092,131	14387,538
	1=sports, 2=politics, 3=lifestyle	4411,963	514,598	,652	8,574	,000	3031,708	5792,219
	number of new articles	415,654	86,600	,365	4,800	,000	183,376	647,932

a. Dependent Variable: views

Πίνακας 4.2: Ο πίνακας που προκύπτει από το μενού Analyze | Regression | Linear του SPSS (2).

7)Να βρεθεί το βέλτιστο μοντέλο γραμμικής παληνδρόμησης με την μέθοδο Stepwise.

Ομοίως με το 1: Analyze|Regression|Linear. Η διαφορά είναι ότι αλλάζουμε το method και βάζουμε Stepwise.

- Χρησιμοποιώντας το πλήρες μοντέλο, να εκτιμηθούν οι συντελεστές της γραμμική του εξίσωσης. Να δοθεί η ερμηνεία των αποτελεσμάτων.

Ο πίνακας 4.1 δείχνει το πλήρες μοντέλο. Οι συντελεστές της γραμμικής εξίσωσης δίνονται στην στήλη B του πίνακα. Άρα η εξίσωση είναι η $Y_{views} = 11819.566 + X_{country} \cdot 457.290 + X_{subject} \cdot 4421.575 + X_{news} \cdot 416.238 + X_{journalist} \cdot 3.461 + X_{yr} \cdot (-426.137)$

Οι συντελεστές δείχνουν κάθε επιπλέον μονάδα του συντελεστή πόσο επηρεάζει το κέρδος μας. Επομένως π.χ. κάθε επιπλέον δημοσιογράφος προσθέτει 3.461 στον ετήσιο αριθμό επισκέψεων της συγκεκριμένης σελίδας κ.λ.π. .

- Χρησιμοποιώντας το πλήρες μοντέλο, να εκτιμηθεί σημειακά και με διάστημα εμπιστοσύνης 99% ο αναμενόμενος επιπρόσθετος ετήσιος αριθμός επισκέψεων, που θα παρουσιάσει μία ελληνική ιστοσελίδα, έναντι μίας όχι ελληνικής με τα ίδια χαρακτηριστικά.

Ανάμεσα σε μία ελληνική σελίδα και ξένη ιστοσελίδα η διαφορά ανάμεσα στην επισκεψιμότητα τους είναι ο συντελεστής country της γραμμικής εξίσωσης (δηλαδή 457.290) υπό την προϋπόθεση ότι οι ιστοσελίδες αυτές παρουσιάζουν τα ίδια χαρακτηριστικά ως προς τις υπόλοιπες μεταβλητές.

Το διάστημα εμπιστοσύνης δίνεται πάλι από τον πίνακα 4.1 στην τελευταία στήλη όπου ο συντελεστής country μπορεί να κυμαίνεται μεταξύ των τιμών -2196.741 και $3111, 321$. Από αυτό το μεγάλο εύρος που έχει ο συντελεστής καταλαβαίνουμε ότι δεν είναι σημα-

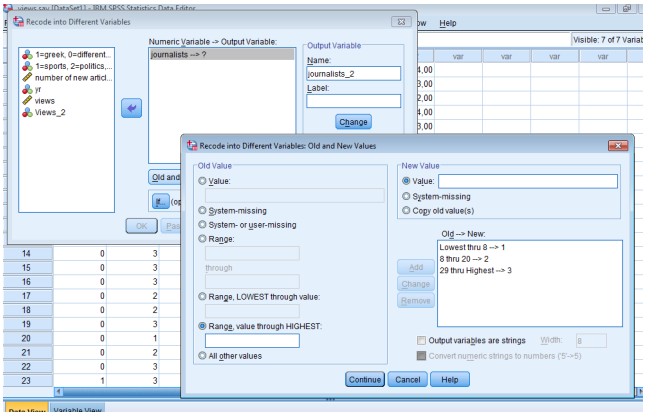
ντικός παράγοντας της εξίσωσης (πράγμα που φαίνεται και από το $p - value$).

5. Μέρος Δ

Δημιουργούμε μία νέα μεταβλητή (journalists_2) που ομαδοποιεί τις ιστοσελίδες ανάλογα με τον αριθμό δημοσιογράφων που απασχολούν ως ακολούθως:
1η ομάδα: ιστοσελίδες με αριθμό δημοσιογράφων μέχρι 8
2η ομάδα: ιστοσελίδες με αριθμό δημοσιογράφων πάνω από 8 μέχρι και 20
3η ομάδα: ιστοσελίδες με αριθμό δημοσιογράφων πάνω από 20

- Εφαρμόζοντας κατάλληλο στατιστικό μοντέλο, να εξετασθεί σε επίπεδο σημαντικότητας 5%, αν ο ετήσιος αριθμός επισκέψεων μίας ιστοσελίδας εξαρτάται από το αν η ιστοσελίδα ανήκει στην 1η, 2η ή 3η ομάδα βάσει του παράγοντα journalists_2 και από τη χώρα προέλευσης. Δώστε την τελική μορφή του μοντέλου στην οποία καταλήξατε και σχολιάστε τα αποτελέσματα.

Για τον ορισμό της νέας μεταβλητής στο SPSS, όπως και πριν, πηγαίνουμε στο μενού: Transform | Recode Into different Variables... (βλ. σχήμα 5.1).



Σχήμα 5.1: Το μενού Transform | Recode Into different Variables... του SPSS.

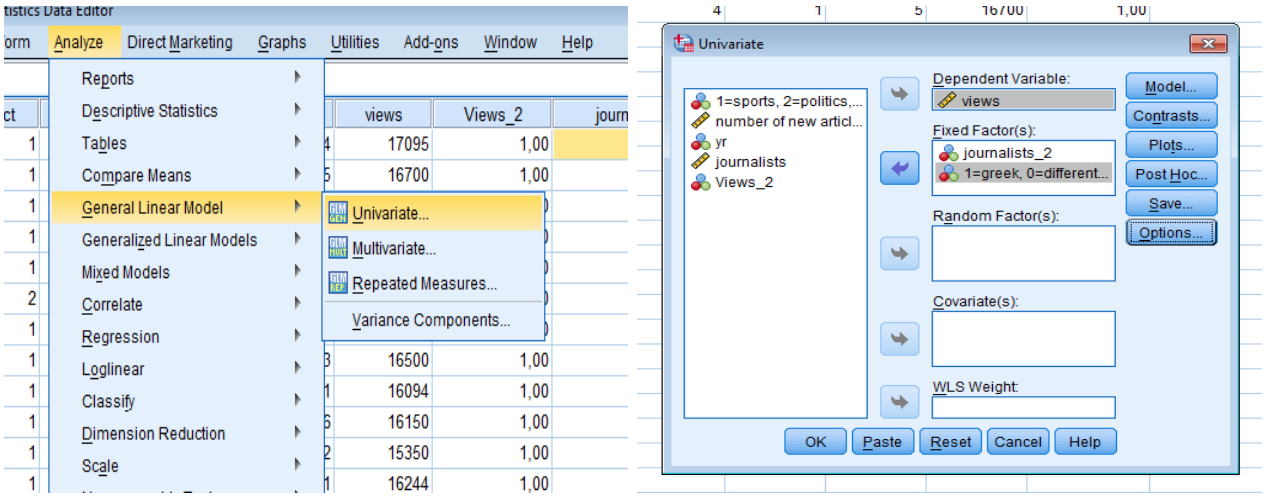
Από την στιγμή που η εξαρτημένη μεταβλητή είναι μία και ποσοτική, ενώ οι εξαρτημένες μεταβλητές είναι ποιοτικές, θα πραγματοποιηθεί ανάλυση διακύμανσης.

Κατ' ουσία είναι έλεγχος σημαντικότητας και ισοδυναμεί με τον ακόλουθο στατιστικό έλεγχο:

$$H_0 : \beta = 0, H_1 : \beta \neq 0 \tag{8}$$

Οι δύο υποθέσεις που έρχονται σε αντιπαράθεση σύμφωνα με την εκφώνηση είναι οι: H_0 και H_1 όπου η H_0 : η εξαρτημένη μεταβλητή είναι ανεξάρτητη από τον παράγοντα και H_1 η εξαρτημένη μεταβλητή είναι εξαρτημένη από τον παράγοντα, άρα και ο παράγοντας είναι σημαντικός.

Για τον έλεγχο της υπόθεσης στο SPSS πηγαίνουμε στο μενού: Analyze|General Linear Model| Univariate (βλ. σχήμα 5.2) και προκύπτει ο πίνακας 5.1.



Σχήμα 5.2: Το μενού Analyze|General Linear Model| Univariate

Tests of Between-Subjects Effects

Dependent Variable: views

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	811494729[a]	5	162298945,9	8,714	,000
Intercept	1,948E+10	1	1,948E+10	1045,837	,000
journalists_2	646450367,5	2	323225183,8	17,354	,000
country	38288944,52	1	38288944,52	2,056	,159
journalists_2 * country	12849158,56	2	6424579,279	,345	,710
Error	838161628,7	45	18625813,97		
Total	2,999E+10	51			
Corrected Total	1649656358	50			

a. R Squared = ,492 (Adjusted R Squared = ,435)

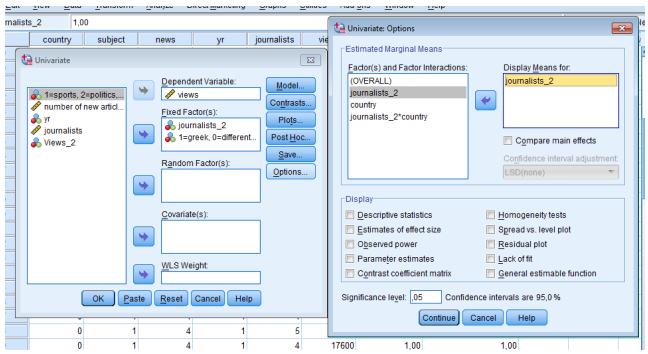
Πίνακας 5.1: Ο πίνακας που προκύπτει από το μενού Analyze|General Linear Model| Univariate του SPSS

Αν το $p - value < 0.05$ τότε αποδεχόμαστε την H_1, δηλαδή αποδεχόμαστε ότι ο παράγοντας είναι σημαντικός.Αν το $p - value > 0.05$ τότε αποδεχόμαστε την H_0. Επομένως από τον πίνακα 5.1 προκύπτει ότι (σε επίπεδο σημαντικότητας 95%) η καινούργια μεταβλητή journalists_2 είναι σημαντικός παράγοντας ενώ η χώρα δεν είναι σημαντικός παράγοντας, επομένως και η αλληλεπίδραση μεταξύ τους δεν είναι σημαντική στατιστική. Συνεπώς το τελικό μοντέλο είναι το ακόλουθο:

$$Y = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \tag{9}$$

- Χρησιμοποιώντας την τελική μορφή του μοντέλου που καταλήξατε στο ερώτημα (1), να δοθούν οι σημειακές εκτιμήσεις και τα διαστήματα εμπιστοσύνης 95% για τους μέσους ετήσιους αριθμούς επισκέψεων των ιστοσελίδων για κάθε μία από τις ομάδες που έχουν σχηματισθεί βάσει της μεταβλητής journalists_2. Σχολιάστε και τα αποτελέσματα.

Για την εύρεση των σημειακών εκτιμήσεων για τους μέσους ετήσιους αριθμούς επισκέψεων στο SPSS, όπως και πριν, πηγαίνουμε στο μενού: Analyze|General Linear Model| Univariate. Μόνο που στην επιλογή Options τώρα μεταφέρουμε τον παράγοντα journalists_2 στο Display Means for (βλ. σχήμα 5.3).



Σχήμα 5.3: Το μενού Analyze|General Linear Model| Univariate του SPSS (2).

journalists_2

Dependent Variable: views

journalists_2	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1,00	17588,500	1137,304	15297,853	19879,147
2,00	23642,067	1364,764	20893,290	26390,843
3,00	26955,115	1135,552	24667,995	29242,235

Πίνακας 5.2: Ο πίνακας που προκύπτει από το μενού Analyze|General Linear Model| Univariate του SPSS (2)

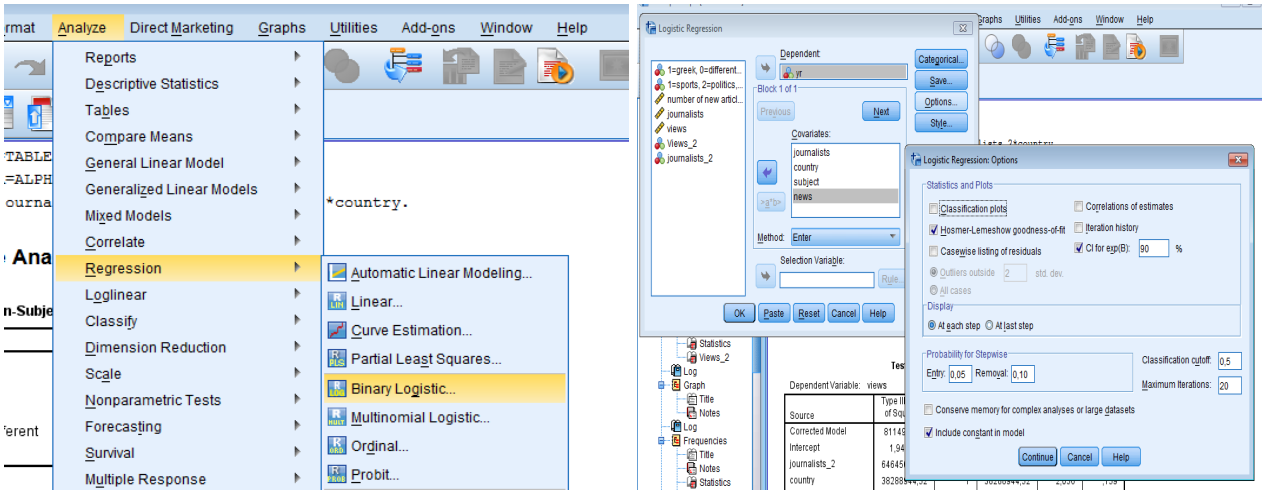
Στον πίνακα 5.2 εμφανίζονται οι μέσοι όροι για κάθε παράγοντα της εξαρτημένης μεταβλητής. Αυτό που παρατηρούμε είναι ότι στην 2 και 3 ομάδα (δηλαδή όσες ιστοσελίδες έχουν μεταξύ 8 έως 20 δημοσιογράφους και όσες ιστοσελίδες έχουν πάνω από 20 δημοσιογράφους) υπάρχει επικάλυψη. Οπότε στατιστικά δεν είναι σίγουρο ότι ο αριθμός των δημοσιογράφων αυξάνει και τον αριθμό της επισκεψιμότητας.

6. Μέρος Ε

- Να εφαρμοσθεί κατάλληλη στατιστική μέθοδος ώστε να διευκρινιστεί το αν οι μεταβλητές journalists, country, subject, news που αντιστοιχούν σε μία ιστοσελίδα είναι επαρκείς πληροφορίες ώστε να μπορούμε να προβλέψουμε την παλαιότητα της συγκεκριμένης ιστοσελίδας. Να βρεθεί το βέλτιστο μοντέλο πρόβλεψης σε επίπεδο σημαντικότητας 10% και να δοθεί η εξίσωση που αντιστοιχεί σε αυτό.

Η εξαρτημένη μεταβλητή (yr) είναι ποιοτική μεταβλητή με 2 επίπεδα. Επομένως ως μέθοδο θα χρησιμοποιήσουμε την λογιστική παλινδρόμηση. Αρχικά θα εφαρμοστεί έλεγχος καλής προσαρμογής του μοντέλου της Λογιστικής Παλινδρόμησης με εξαρτημένη τη μεταβλητή Yr και ανεξάρτητες όλες τις υπόλοιπες μεταβλητές. Ο έλεγχος Hosmer & Lemeshow test πραγματοποιείται έτσι ώστε να επιβεβαιωθεί ότι το σύνολο των παραγόντων που έχουν καταγραφεί ως υποψήφιοι να επιδρούν πάνω στο τελικό αποτέλεσμα, έχουν καλή προσαρμογή στα δεδομένα του προβλήματος.

Για τον υπολογισμό της λογιστικής παλινδρόμησης καθώς και του ελέγχου της στο SPSS, πηγαίνουμε στο μενού: Analyze | Regresion | Binary Logistic (βλ. σχήμα 6.1) και τα αποτελέσματα που μας ενδιαφέρουν φαίνονται στον πίνακα 6.1 και 6.2.



Σχήμα 6.1: Το μενού Analyze | Regresion | Binary Logistic

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	14,128	8	,078

Πίνακας 6.1: Ο πίνακας που προκύπτει από το μενού Analyze | Regresion | Binary Logistic του SPSS

Η τιμή (από τον πίνακα 6.1) $p - value = 0.078 > 0.10$, συνεπώς σε επίπεδο σημαντικότητας 10%, δεν απορρίπτουμε την μηδενική υπόθεση της καλής προσαρμογής, που ση-

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	90% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	journalists	-,205	,065	9,951	1	,002	,814	,732	,906
	country	,982	,976	1,013	1	,314	2,670	,536	13,301
	subject	1,430	,638	5,018	1	,025	4,178	1,462	11,938
	news	,049	,096	,256	1	,613	1,050	,896	1,229
	Constant	,673	1,076	,391	1	,532	1,959		

a. Variable(s) entered on step 1: journalists, country, subject, news.

Πίνακας 6.2: Ο πίνακας που προκύπτει από το μενού Analyze | Regresion | Binary Logistic του SPSS(2)

μαίνει ότι το μοντέλο που εφαρμόσαμε παρουσιάζει συνολικά καλή προσαρμογή στα δεδομένα.

Στον πίνακα 6.2 φαίνεται πιο παράγοντες επηρεάζουν σε σημαντικό βαθμό την εξαρτημένη μεταβλητή. Οι παράγοντες των οποίων η τιμή $p - value$ είναι μεγαλύτερη από το επίπεδο σημαντικότητας 10% δεν επηρεάζουν σε σημαντικό βαθμό την εξαρτημένη μεταβλητή. Δηλαδή η μόνη σημαντική μεταβλητή είναι ο αριθμός των δημοσιογράφων.

σελ.51 -53 η απάντηση

- Χρησιμοποιώντας το βέλτιστο μοντέλο, να προβλεφθεί η παλαιότητα μίας ελληνικής ιστοσελίδας για την οποία γνωρίζουμε ότι απασχολεί 12 δημοσιογράφους, πραγματεύεται θέματα αθλητικής επικαιρότητας και στην οποία αναρτώνται ημερησίως 15 νέα άρθρα.

σελ.56 η απάντηση

Πρέπει πρώτα να συμπληρώσουμε μία επιλέον γραμμή με αυτά που θέλουμε. Analyze|Regresion| Binary Logistic. Στην επιλογή Save στο Predicted Values τικάρουμε το Probabilities. Δίνει απ' ευθείας την πιθανότητα και συγκεκριμένα η πιθανότητα y=1

Αναφορές

[1] Wikipedia. Median — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Median. [Πρόσβαση στις 23 Ιουλίου 2014].

[2] Wikipedia. Mode (statistics) — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Mode_%28statistics%29. [Πρόσβαση στις 23 Ιουλίου 2014].

[3] Wikipedia. Percentile — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Percentile. [Πρόσβαση στις 23 Ιουλίου 2014].

[4] Wikipedia. Διακύμανση — Wikipedia, the free encyclopedia. http://el.wikipedia.

org/wiki/%CE%94%CE%B9%CE%B1%CE%BA%CF%8D%CE%BC%CE%B1%CE%BD%CF%83%CE%B7. [Πρόσβαση στις 23 Ιουλίου 2014].

[5] Wikipedia. Μέσος όρος — Wikipedia, the free encyclopedia. http://el.wikipedia.org/wiki/%CE%9C%CE%AD%CF%83%CE%BF%CF%82_%CF%8C%CF%81%CE%BF%CF%82. [Πρόσβαση στις 23 Ιουλίου 2014].

[6] Θ.Κ. Κωνσταντινίδης Στυλιανός Κ. Τσίπος. Βασικές Αρχές Βιοστατιστικής - Εφαρμογές με χρήση του spss, 2010.

[7] Ιωάννης Τριανταφύλλου. Σημειώσεις "Πιθανότητες - Στατιστική". Σημειώσεις μαθήματος, 2012.

bin/]ergasia.mintedcmdbin/]ergasia.mintedmd5bin/]ergasia.pygbin/]ergasia.out.pyg