



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΠΛΗΡΟΦΟΡΙΚΗ»**

Πιθανότητες και Στατιστική Εργασία Εξαμήνου

Αναγνωστόπουλος Βασίλης - Θάνος

ΑΘΗΝΑ, 2014

© Αθήνα, 2014 Αναγνωστόπουλος Βασίλης - Θάνος

Το κείμενο αυτό έχει γραφτεί σε \LaTeX .

Αυτό το κείμενο διανέμεται σύμφωνα με τους όρους της άδειας Creative Commons Attribution - ShareAlike Unported 3.0.

Εν συντομία: Είστε ελεύθεροι να διανέμετε και να τροποποιήσετε αυτό το κείμενο εφόσον αναφέρετε τον δημιουργό του και διατηρήσετε την ίδια άδεια χρήσης.

Το παρόν έγγραφο διανέμεται με την ελπίδα ότι θα είναι χρήσιμο, αλλά χωρίς καμία εγγύηση, χωρίς ακόμη και την έμμεση εγγύηση εμπορευσιμότητας ή καταλληλότητας για κάποιο συγκεκριμένο σκοπό.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευτεί ότι αντιπροσωπεύουν το Πανεπιστήμιο Πειραιώς

Περιεχόμενα

| | | |
|----------|---|-----------|
| 1 | Εισαγωγή - Αντικείμενο της άσκησης | 1 |
| 2 | Μέρος Α | 1 |
| 3 | Μέρος Β | 8 |
| 4 | Μέρος Γ | 13 |
| 5 | Μέρος Δ | 14 |
| 6 | Μέρος Ε | 15 |
| | Βιβλιογραφία | 15 |

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 2 παρατήρηση με τιμές (0,1,12,1,22,35350)

2. Μέρος Α

- Να υπολογισθεί η 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 n 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]:

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

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

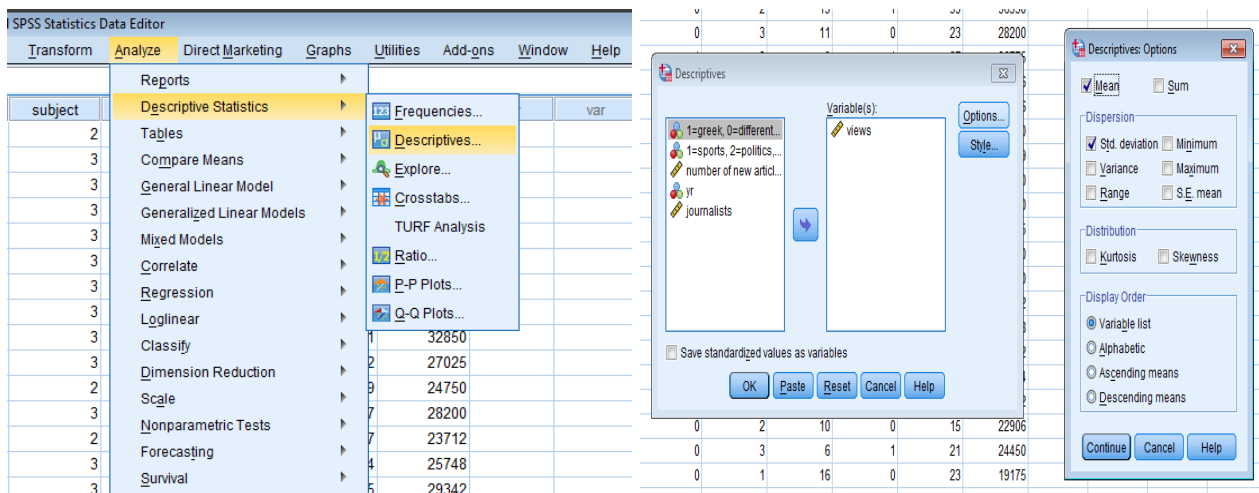
και συμβολίζεται με σ . Ο γενικός τύπος της απόκλισης είναι [4]:

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

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 βαθμολογιών 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 n 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].

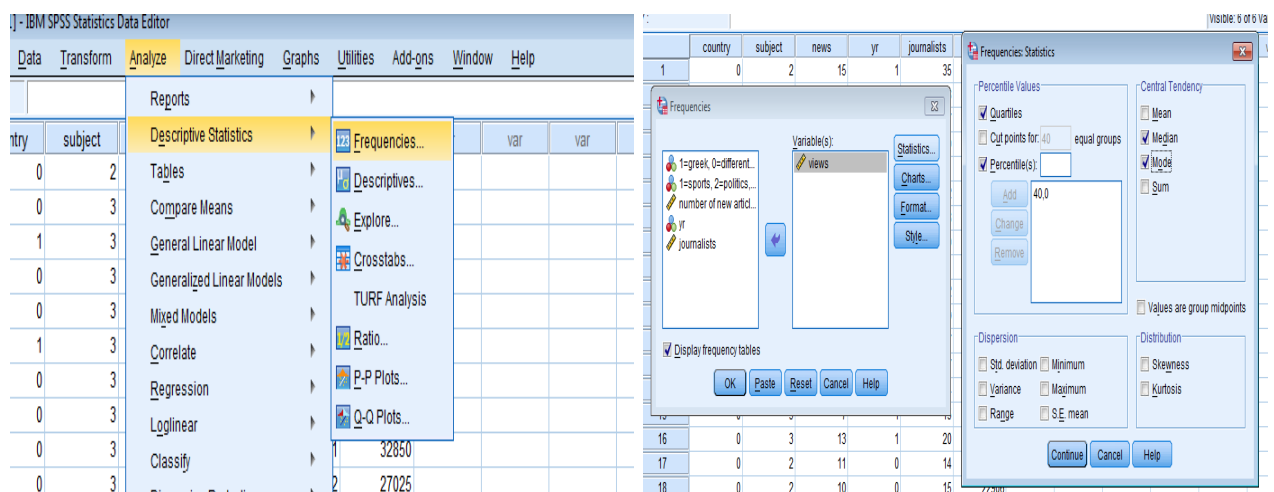
Κορυφή είναι η τιμή που εμφανίζεται πιο συχνά σε ένα σύνολο δεδομένων [2].

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 κατηγορίες ανάλογα με τον ετήσιο αριθμό επισκέψεων τους ως εξής:

1η ομάδα: ιστοσελίδες με ετήσιο αριθμό επισκέψεων μέχρι 22000

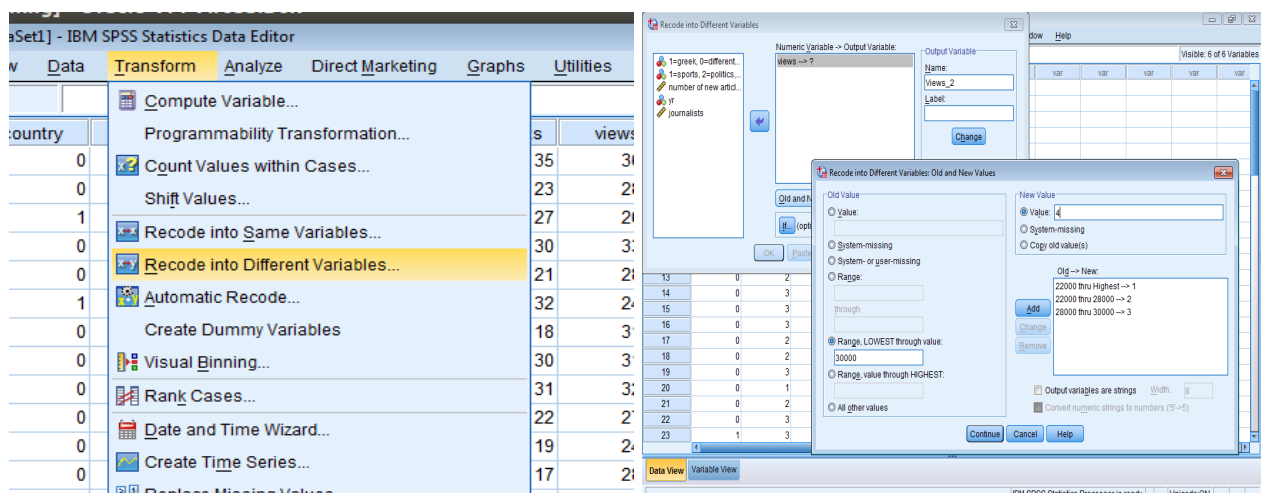
2η ομάδα: ιστοσελίδες με ετήσιο αριθμό επισκέψεων πάνω από 22000 και μέχρι 28000

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).



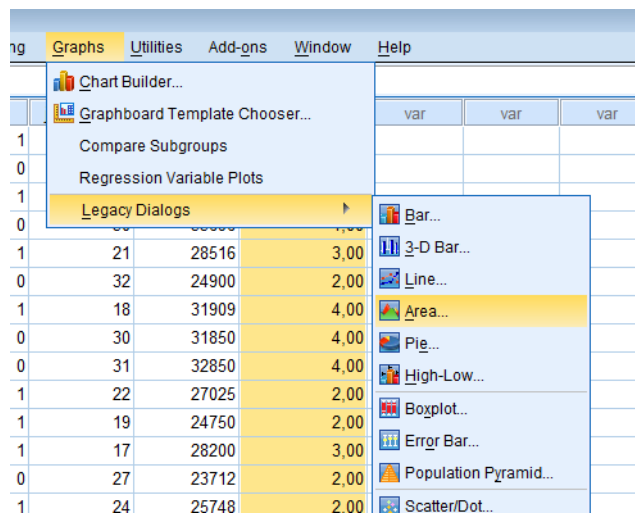
Σχήμα 2.3: Το μενού Transform | Recode Into different Variables... του SPSS

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.

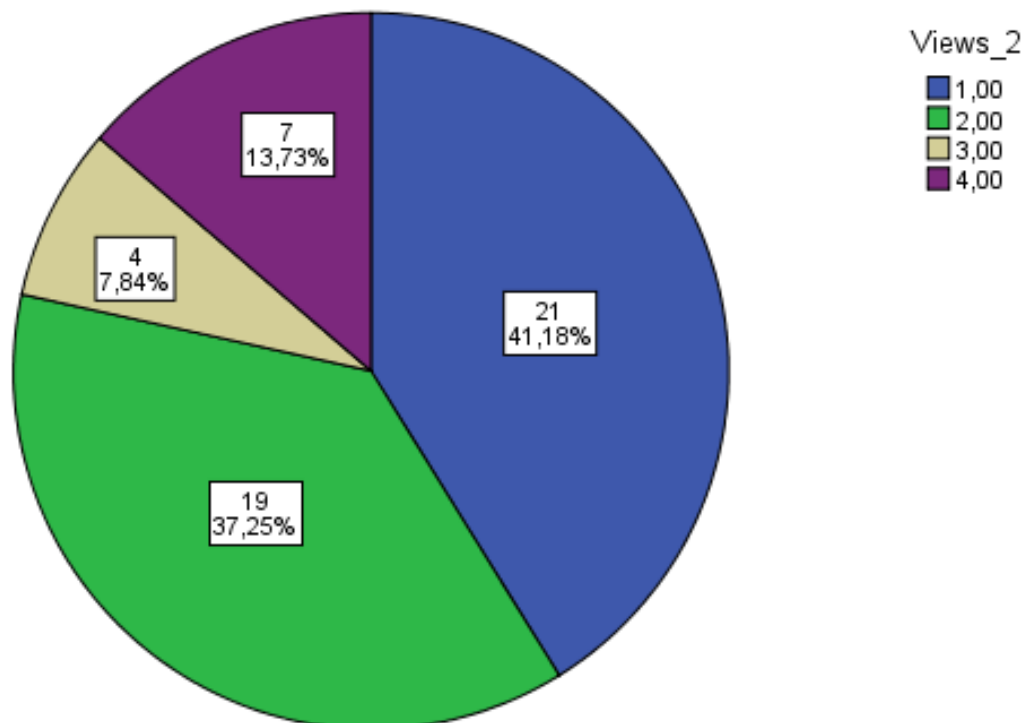
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|-------|------|-----------|---------|---------------|--------------------|
| Valid | 1.00 | 21 | 41.2 | 41.2 | 41.2 |
| | 2.00 | 19 | 37.3 | 37.3 | 78.4 |
| | 3.00 | 4 | 7.8 | 7.8 | 86.3 |
| | 4.00 | 7 | 13.7 | 13.7 | 100.0 |
| Total |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

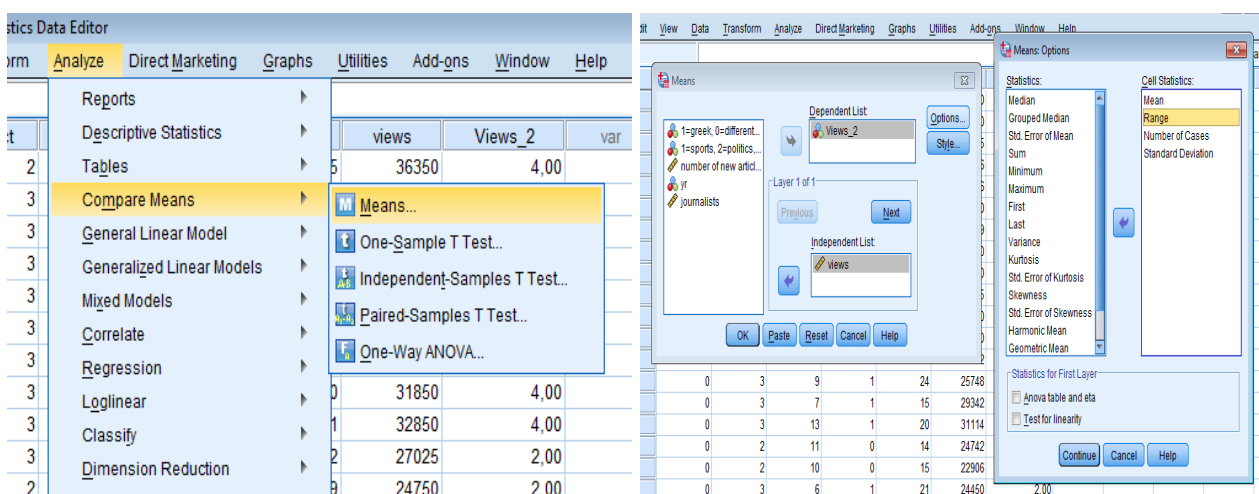
Μερικά από τα βασικότερα μέτρα διασποράς ή μεταβλητότητας είναι:

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

$$R = t_{max} - t_{min} \quad (3)$$

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.

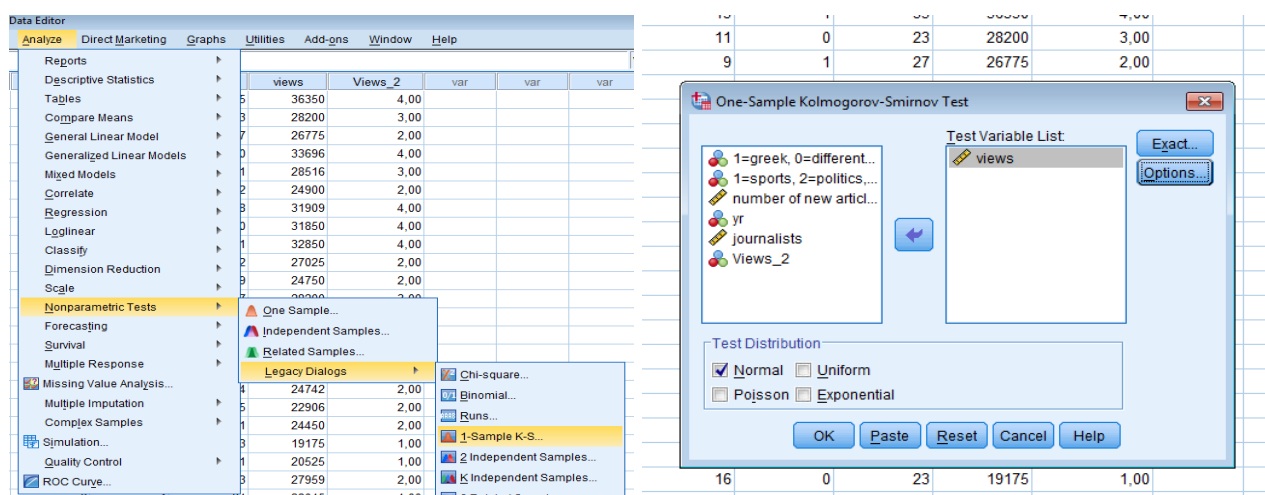
Οι δύο υποθέσεις που έρχονται σε αντίθεση σύμφωνα με την εκφώνηση είναι οι ακόλουθες:

$$H_0 : \mu_g = \mu_f, H_1 : \mu_g \neq \mu_f \quad (4)$$

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

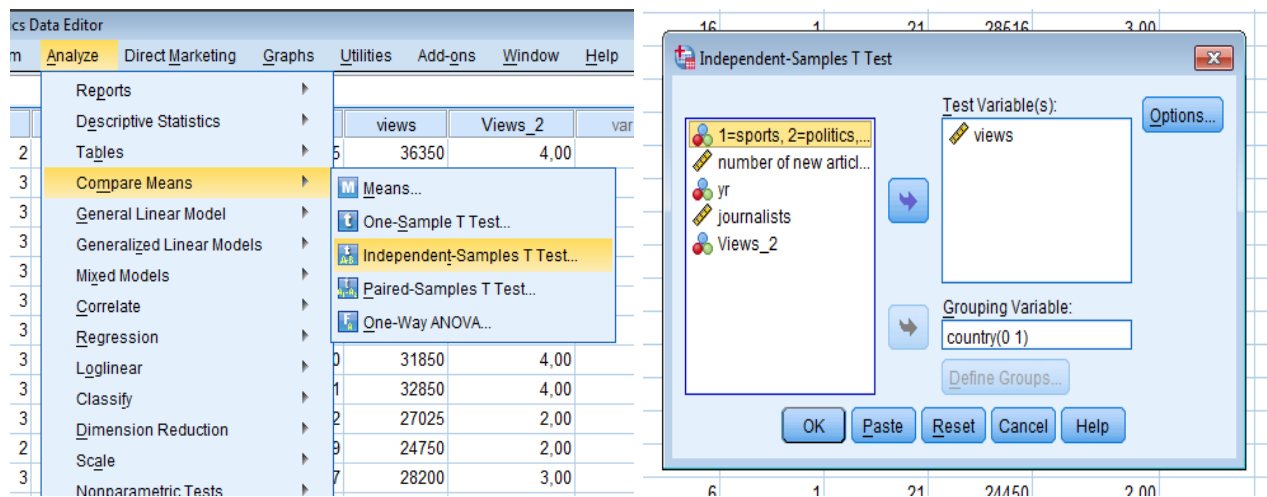
| | | views |
|--------------------------|----------------|----------|
| N | | 51 |
| Normal Parameters | Mean | 23571,14 |
| | Std. Deviation | 5743,964 |
| Most Extreme Differences | Absolute | ,095 |
| | Positive | ,095 |
| | Negative | -,068 |
| Test Statistic | | ,095 |
| Asymp. Sig. (2-tailed) | | ,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 | | | | | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference |
| 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:

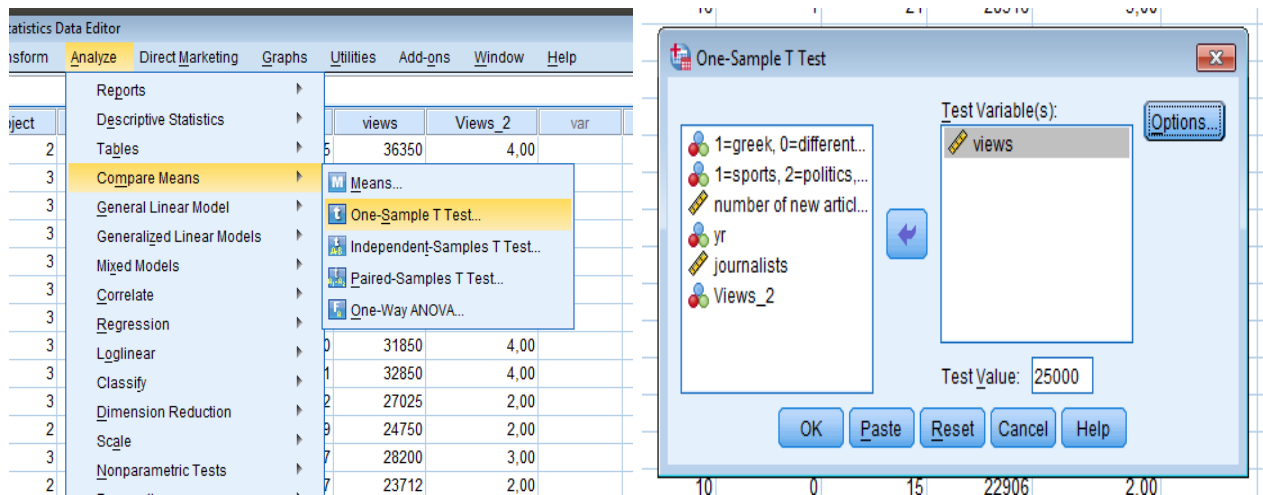
$$H_0 : \mu_{<2} = \mu_{>2}, H_1 : \mu_{<2} \neq \mu_{>2} \quad (5)$$

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 | | | | | | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the 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 |



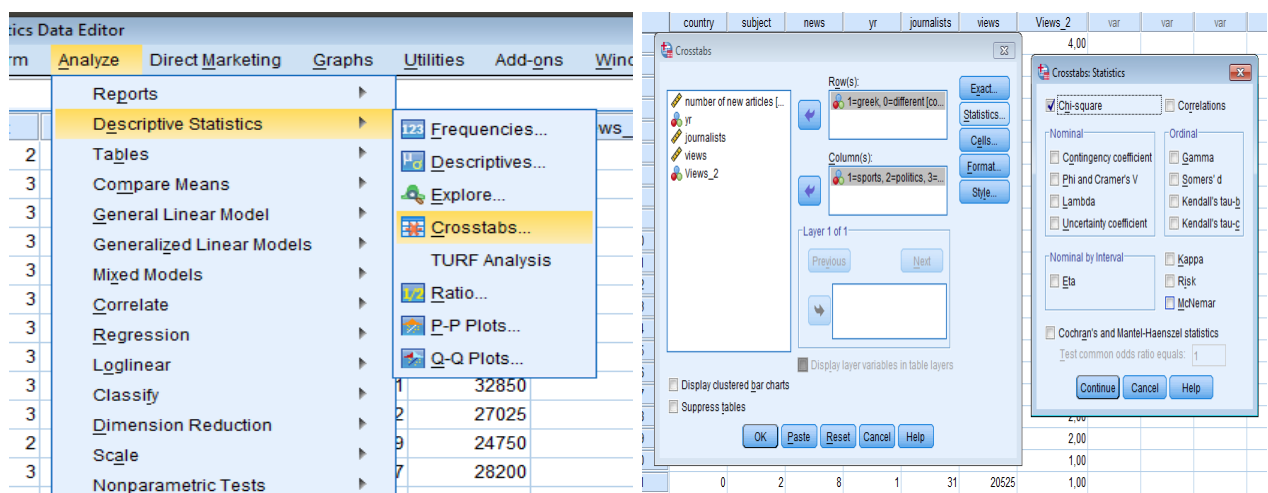
Σχήμα 3.3: Το μενού Analyze | Compare means | Independent samples T-test του SPSS

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 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 | | | Total |
|----------------------|---|--|-----------------------------------|--------|--------|--------|
| | | | 1 | 2 | 3 | |
| 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.
- Χρησιμοποιώντας το πλήρες μοντέλο, να εκτιμηθούν οι συντελεστές της γραμμικής του εξίσωσης. Να δοθεί η ερμηνεία των αποτελεσμάτων.
- 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.

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.

- 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.

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.

- 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.

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