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 XeLaTeX.
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

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

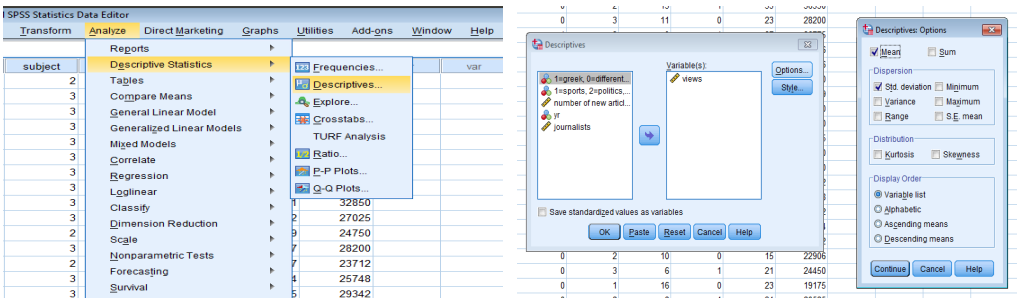
- Να υπολογισθεί ή μέση τιμή και η τυπική απόκλιση των ετήσιων αριθμών επισκέψεων των ιστοσελίδων. Να δοθεί η ερμηνεία των αποτελεσμάτων.

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



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

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.

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 [6].

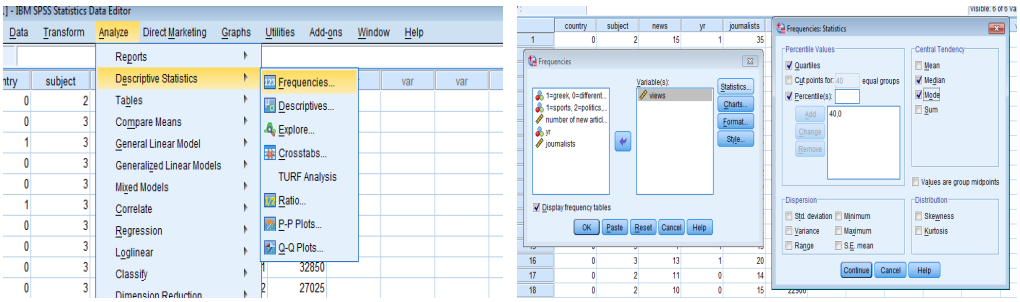
- Να υπολογισθεί η διάμεσος, τα τεταρτημόρια, το 40% ποσοστημόριο και η κορυφή των ετησίων αριθμών επισκέψεων των ιστοσελίδων. Να δοθεί η ερμηνεία των αποτελεσμάτων.

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 [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 Statis-



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

tics/Frequencies (βλ. σχήμα 2.3) 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 .

- 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

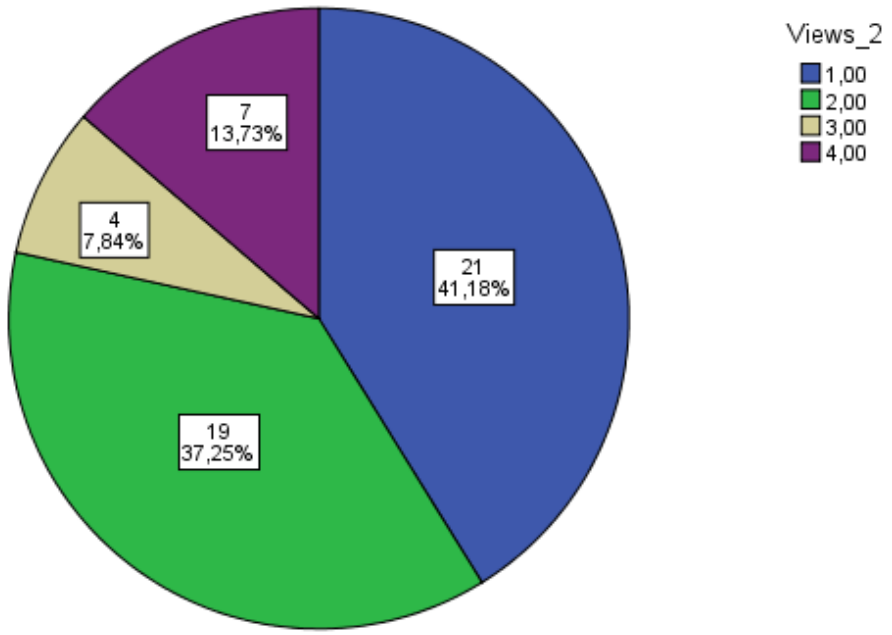
  – Να κατασκευασθεί ο πίνακας συχνοτήτων και το κυκλικό διάγραμμα βάσει της νέας μεταβλητής

  – Τί ποσοστό των ιστοσελίδων ανήκουν στην 3η ομάδα;

Για τον 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.



Σχήμα 2.3: Το μενού Analyze/Descriptive Statistics/Frequencies του SPSS

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

3. Μέρος Β

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

N	Valid	51
	Missing	0
Median		2.000
Mode		1.00
std. Deviation		1.02785
Variance		1.056
Range		3.000
Percentiles	25	1.00
Percentiles	50	2.00
Percentiles	75	3.00

Πίνακας 2.4: Η διάμεσος, τα τεταρτημόρια, το ποσοστημόριο και η κορυφή των ετήσιων αριθμών επισκέψεων των ιστοσελίδων.

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

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

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. ελληνικής με τα ίδια χαρακτηριστικά.

- Χρησιμοποιώντας το πλήρες μοντέλο, να εκτιμηθούν οι συντελεστές της γραμμική 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] Ιωάννης Τριανταφύλλου. Σημειώσεις "Πιθανότητες - Στατιστική". Σημειώσεις μαθήματος, 2012.

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