ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΔΗΜΟΓΡΑΦΙΚΟΙ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΙΚΟΙ ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΣΧΕΤΙΖΟΝΤΑΙ ΜΕ ΤΗΝ ΕΞΑΠΛΩΣΗ ΤΟΥ ΚΟΡΩΝΟΪΟΥ ΣΤΗ ΝΟΤΙΑ ΚΟΡΕΑ. ΜΙΑ ΜΕΛΕΤΗ ΤΗΣ KCDC DATA BASE.

ΚΟΜΙΑΝΟΥ-ΓΑΛΑΝΟΠΟΥΛΟΥ ΦΩΤΕΙΝΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΠΟΥ ΥΠΟΒΛΗΘΗΚΕ ΣΤΟ ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΤΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΕΙΡΑΙΩΣ ΩΣ ΜΕΡΟΣ ΤΩΝ ΑΠΑΙΤΗΣΕΩΝ ΓΙΑ ΤΗΝ ΑΠΟΚΤΗΣΗ ΤΟΥ ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ.

ΠΕΙΡΑΙΑΣ
ΝΟΕΜΒΡΙΟΣ 2021

UNIVERSITY OF PIRAEUS

SCHOOL OF FINANCE AND STATISTICS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

DEMOGRAPHIC AND ENVIRONMENTAL FACTORS RELATED TO THE SPREAD OF COVID-19 IN SOUTH KOREA. A STUDY OF THE KCDC DATA BASE

by

KOMIANOU-GALANOPOULOU FOTEINI

MSC DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE OF UNIVERSITY OF PIRAEUS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN APPLIED STATISTICS.

PIRAEUS
NOVEMBER 2021

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίστηκε από τη ΓΣΕΣ του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ΄ αριθμ…… συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της επιτροπής ήταν:

1. Τζαβελάς Γεώργιος, Αναπληρωτής Καθηγητής (επιβλέπων)
2. Βερροπούλου Γεωργία, Καθηγήτρια
3. Λαβδανίτη Μαρία, Διεθνές Πανεπιστήμιο της Ελλάδος, Καθηγήτρια

Η έγκριση της Διπλωματικής Εργασίας από το τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω πολύ τον επιβλέποντα καθηγητή μου κ. Γεώργιο Τζαβελά για την βοήθειά του καθώς επίσης και για την επιμονή και υπομονή που είχε μαζί μου όλο αυτό το διάστημα.

ΠΕΡΙΛΗΨΗ

Ο κορωνοϊός είναι μια ασθένεια που εμφανίστηκε στα τέλη Δεκεμβρίου 2019 στην πόλη Wuhan της Κίνας η οποία προκαλεί σοβαρές αναπνευστικές λοιμώξεις στον ανθρώπινο οργανισμό. Η ταχεία μετάδοσή του είχε ως αποτέλεσμα να κηρυχθεί ο ιός αυτός ως πανδημία σε όλο τον κόσμο.

Βασικοί στόχοι της συγκεκριμένης διπλωματικής εργασίας είναι να εντοπιστούν οι δημογραφικοί παράγοντες που συμβάλλουν στην εξάπλωση του COVID-19 στους πολίτες της Νοτίου Κορέας καθώς και πως οι περιβαλλοντικοί παράγοντες της χώρας συσχετίζονται με τη συγκεκριμένη ασθένεια.

Τα στοιχεία που χρησιμοποιήθηκαν για την ανάλυση της διπλωματικής εργασίας πάρθηκαν από τον Οργανισμό Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA) ο οποίος συγκεντρώνει στοιχεία από τις τοπικές κυβερνήσεις. Η αφετηρία της έρευνας του συγκεκριμένου οργανισμού είναι στις 20 Ιανουαρίου 2020 όπου και εμφανίστηκε το πρώτο κρούσμα και η οποία συνεχίζει να συλλέγει πληροφορίες έως και σήμερα.

Οι στατιστικές μέθοδοι που χρησιμοποιήθηκαν στην έρευνά μας είναι η μονοδιάστατη και δισδιάστατη περιγραφή ανάλυση, ο έλεγχος χ^2 του Pearson, η πολλαπλή λογιστική παλινδρόμηση, το μοντέλο αναλογικού κινδύνου του Cox, ο συντελεστής συσχέτισης του Pearson όπως και η μέθοδος Box και Jenkins της ανάλυσης χρονολογικών σειρών.

Η ανάλυση δεδομένων δείχνει ότι οι παράγοντες που σχετίζονται με την εξάπλωση του κορωνοϊού είναι το φύλο, η ηλικία και η επαρχία. Από τους περιβαλλοντικούς παράγοντες οι οποίοι μελετήθηκαν στην πόλη της Seoul, η μέση ημερήσια θερμοκρασία και η μέση σχετική υγρασία σχετίζονται επίσης με τον αριθμό κρουσμάτων. Τέλος, προτείνεται ένα μοντέλο χρονοσειρών το οποίο μπορεί να προβλέπει τον αριθμό των κρουσμάτων.

ABSTRACT

Coronavirus is a disease that appeared in late December 2019 in Wuhan city of China wand it causes severe respiratory infections to people. Because of its rapid spread, has been declared as a pandemic around the world.

The main objectives of this thesis are to detect the demographic factors contributing to the spread of COVID-19 to South Korean citizens and how the country's environmental factors are related to this disease.

The data used to analyze the this thesis was obtained by the South Korea Disease Control and Prevention Agency (KCDA) which collects data from local governments. The KCDA began its investigation on 20 January 2020 when the first case occurred and it continues to collect information to this day.

The statistical methods used for our research, are the one-dimensional and two-dimensional descriptive statistics, Pearson test of X^2, Multinomial Logistic Regression, the COX Regression, Pearson Bivariate Correlations as well as the Box and Jenkins method of Sequence Charts.

Data analysis reveals that the gender, the age and the province are the most important factors related to the spread of the coronavirus. The environmental factors are studied at the Seoul Province and it shown that the mean daily temperature and the mean relative humidity are related to the spread of the coronavirus. Finally a times series model is proposed which can predict the number of new coronavirus cases.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Πινάκων
Κατάλογος Σχημάτων

ΚΕΦΑΛΑΙΟ 1
Εισαγωγή……1

1.1 Κορωνοϊός……1

1.1.1 Ιστορία του Κορωνοϊού…….……..1
1.1.2 Σύγκριση Κορωνοϊού με ιό Γρίπης (H1N1)…….1
1.1.3 Συμπτώματα και μετάδοση στον ανθρώπινο οργανισμό……...2
1.1.4 Στατιστικά στοιχεία και αντιμετώπιση……..……..2
1.1.5 Εμφάνιση Κορωνοϊού στη Νότια Κορέα……..3

1.2 Μελέτη του Κέντρου Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDC) …….....4

1.2.1 Ιστορία του Οργανισμού Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA)……..4
1.2.2 Τομείς δραστηριότητας του Οργανισμού Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA)…….....4
1.2.3 Αναφορά στην μελέτη του Οργανισμού Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA) για τον Κορωνοϊό……....4

1.3 Δεδομένα……..……5

1.3.1 Στοιχεία της παρούσας εργασίας…….……5

1.4 Μεθοδολογία……..……..6

1.4.1 Έλεγχος X^2……..……6
1.4.2 Έλεγχος Kruskal-Wallis για *κ* ανεξάρτητα δείγματα……7
1.4.3 Πολλαπλή λογιστική Παλινδρόμηση……8
1.4.4 Καμπύλες ROC……...9
1.4.5 Πίνακες Επιβίωσης με τη μέθοδο Kaplan-Meier……..10
1.4.6 Το μοντέλο αναλογικού κινδύνου (PH) του Cox……10
1.4.7 Συντελεστής συσχέτισης του Pearson……11
1.4.8 Ανάλυση Χρονολογικών Σειρών…….11
1.4.9 Μεθοδολογία του Box και Jenkins…….12

ΚΕΦΑΛΑΙΟ 2

2.1 Περιγραφική Ανάλυση……14

2.1.1 Δημογραφικοί Παράγοντες……..……..14

2.1.1.1Φύλο…….14
2.1.1.2 Ηλικία……15
2.1.1.3 Τόπος Κατοικίας……17
2.1.1.4 Επαρχία Προέλευσης……18
2.1.1.5 Χώρα Προέλευσης…….……20
2.1.1.6 Μέρος εμφάνισης του Κορωνοϊού……21

2.1.2 Περιβαλλοντικοί Παράγοντες……...22

2.1.2.1 Μέσος όρος Θερμοκρασίας……..……22
2.1.2.2 Ελάχιστη Θερμοκρασία……...……23
2.1.2.3 Μέγιστη Θερμοκρασία……..……24
2.1.2.4 Βροχόπτωση…..……..……...25
2.1.2.5 Μέγιστη ταχύτητα Ανέμου…….……..26
2.1.2.6 Κατεύθυνση Ανέμου……..27
2.1.2.7 Μέση σχετική Υγρασία……..……...28

2.2 Κορωνοϊός και άλλες Ασθένειες……..……...29

2.3 Χρονολογική σύγκριση του Κορωνοϊού……...30

2.3.1 Χρονολογική σύγκριση με βάση την Ηλικία…….....31
2.3.2 Χρονολογική σύγκριση με βάση το Φύλο……32
2.3.3 Χρονολογική σύγκριση με βάση την Επαρχία…….....32

ΚΕΦΑΛΑΙΟ 3

3.1 Εισαγωγή……34

3.1.1 Συσχέτιση με τη μεταβλητή ‘‘Φύλο’’……..34
3.1.2 Συσχέτιση με τη μεταβλητή ‘‘Ηλικία’’……35
3.1.3 Συσχέτιση με τη μεταβλητή ‘‘Επαρχία’’…….37

3.2 Λογιστική Παλινδρόμηση με τις μεταβλητές ‘‘Φύλο’’, ‘‘Ηλικία’’ και ‘‘Επαρχία’’……39

3.3 Πίνακες Επιβίωσης……..47

3.3.1 Διάγραμμα Επιβίωσης για το σύνολο των ατόμων……..47

3.3.1.1 Διάγραμμα Επιβίωσης για το σύνολο των ατόμων με την μέθοδο Kaplan-Meier……..48
3.3.2 Διάγραμμα επιβίωσης για τη μεταβλητή ‘‘Φύλο’’..……..48
3.3.2.1 Διάγραμμα Επιβίωσης για τη μεταβλητή ‘‘Φύλο’’ με την μέθοδο Kaplan-Meier..……...49
3.3.3 Διάγραμμα επιβίωσης για τη μεταβλητή ‘‘Ηλικία’’…….49
3.3.3.1 Διάγραμμα Επιβίωσης για τη μεταβλητή ‘‘Ηλικία’’ με την μέθοδο Kaplan-Meier…..……...50
3.3.4 Διάγραμμα επιβίωσης για τη μεταβλητή ‘‘Επαρχία’’..……50
3.3.4.1 Διάγραμμα Επιβίωσης για τη μεταβλητή ‘‘Επαρχία’’ με την μέθοδο Kaplan-Meier……..51

3.4 Το μοντέλο Αναλογικού κινδύνου του Cox με τις μεταβλητές ‘‘Φύλο’’, ‘‘Ηλικία’’ και ‘‘Επαρχία’’……...51

ΚΕΦΑΛΑΙΟ 4

4.1 Εισαγωγή…..……55

4.1.1 Συσχέτιση με τις μεταβλητές ‘‘Μέσος όρος θερμοκρασίας’’, ‘‘Μέση σχετική Υγρασία’’ και ‘‘Κρούσματα’’……55

4.2 Διαγραμματική απεικόνιση των χρονοσειρών ‘‘Μέσος όρος θερμοκρασίας’’, ‘‘Μέση σχετική Υγρασία’’ και ‘‘Κρούσματα’’……..57

4.2.1 Ανάλυση της χρονοσειράς ‘‘Κρούσματα’’…….59
4.2.2 Πρόβλεψη των μελλοντικών τιμών της χρονοσειράς ‘‘Κρούσματα’’……65

Σύνοψη Αποτελεσμάτων……67

Παράρτημα…….68

1 Δεδομένα της παρούσας εργασίας……68

2 Πίνακες συχνοτήτων για τον τόπο κατοικίας και το μέρος εμφάνισης του κορωνοϊού για το 2^{ο} κεφάλαιο……...73

Βιβλιογραφία……..80

Κατάλογος Πινάκων

2.1 Περιγραφή του Φύλου……14

2.2.Πειγραφή των κατηγοριών Ηλικίας……15

2.3.Περιγραφικά μέτρα για την Ηλικία…….16

2.4.Περιγραφή των διάφορων Επαρχιών……..18

2.5.Ρυθμός μεταβολής Κρουσμάτων…….19

2.6.Περιγραφή των διάφορων Χωρών……..……20

2.7.Περιγραφή σημείων εμφάνισης της ασθένειας……21

2.8.Περιγραφικά μέτρα της μέσης Θερμοκρασίας……22

2.9.Περιγραφικά μέτρα ελάχιστης Θερμοκρασίας……23

2.10.Περιγραφικά μέτρα μέγιστης Θερμοκρασίας……24

2.11. Περιγραφικά μέτρα Βροχόπτωσης……25

2.12. Περιγραφικά μέτρα μέγιστης ταχύτητας Ανέμου…….26

2.13. Περιγραφικά μέτρα κατεύθυνσης Ανέμου……27

2.14. Περιγραφικά μέτρα μέσης Υγρασίας……28

2.15. Μέσο Ημερήσιο Ποσοστό των ασθενειών, 2016-2020…….29

2.16.Ποσοστά χρονολογικής σύγκρισης του Κορωνοϊού……...30

2.17. Ποσοστά χρονολογικής σύγκρισης ανά Ηλικία…….31

2.18. Ποσοστά χρονολογικής σύγκρισης ανά Φύλο……...32

2.19. Ποσοστά χρονολογικής σύγκρισης ανά Επαρχία……..33

3.1.Πίνακας συνάφειας ανά Φύλο……..34

3.2.Έλεγχος χ^2 του Pearson ανά Φύλο……34

3.3.Τα σκορ ανά Ηλικία……..36

3.4.Έλεγχος Kruskal-Wallis ανά Ηλικία…….36

3.5.Πίνακας συνάφειας ανά Επαρχία……..37

3.6.Έλεγχος Fisher’s ανά Επαρχία……..38

3.7. Κωδικοποίηση μεταβλητής απόκρισης……..39

3.8. Κωδικοποίηση ανεξάρτητων μεταβλητών……..40

3.9. Ποσοστό μεταβλητότητας του μοντέλου……40

3.10. Έλεγχος καλής προσαρμογής……40

3.11.Πίνακας Ταξινόμησης……41

3.12.Διαχωρισμός θετικών και αρνητικών περιπτώσεων……...41

3.13.Πίνακας αξιολόγησης της περιοχής κάτω από την καμπύλη……..42

3.14.Πίνακας επιλογής σημείου αποκοπής……43

3.15. Κωδικοποίηση μεταβλητής απόκρισης……..44

3.16. Κωδικοποίηση ανεξάρτητων μεταβλητών……44

3.17. Ποσοστό μεταβλητότητας του μοντέλου……...44

3.18. Έλεγχος καλής προσαρμογής……45

3.19.Πίνακας Ταξινόμησης……45

3.20.Εκτίμηση των παραμέτρων του μοντέλου……..46

3.21.Γενικές πληροφορίες δεδομένων……52

3.22.Έλεγχος σημαντικότητας του μοντέλου……..52

3.23.Εκτίμηση παραμέτρων του μοντέλου……53

4.1.Συσχέτιση κρουσμάτων με μέση Θερμοκρασία και μέση Υγρασία.……55

4.2.Συσχέτιση 10ης ημέρας κρουσμάτων με μέση Υγρασία και μέση Θερμοκρασία.56

4.3.Αυτοσυσχετίσεις Κρουσμάτων……60

4.4.Μερικές Αυτοσυσχετίσεις Κρουσμάτων……..61

4.5.Στατιστική συνάρτηση του μοντέλου ARIMA……62

4.6. Παράμετροι του μοντέλου ARIMA.……63

4.7. Αυτοσυσχετίσεις Κρουσμάτων……64

4.8.Μελλοντικές τιμές Κρουσμάτων……...65

Κατάλογος Σχημάτων

2.1. Ποσοστά ασθενών ανά Φύλο……14

2.2. Διακύμανση της ασθένειας με βάση την Ηλικία……..16

2.3. Κατάταξη εμφάνισης της ασθένειας ανά Πόλη……...17

2.4.Κατάταξη των ασθενών ανά Επαρχία……...18

2.5.Ρυθμός μεταβολής κρουσμάτων ανά εβδομάδα……19

2.6.Ποσοστά ασθενών ανά Χώρα……...20

2.7.Κατάταξη σημείων εμφάνισης της ασθένειας……...21

2.8.Διακύμανση της μέσης Θερμοκρασίας…….22

2.9.Διακύμανση ελάχιστης Θερμοκρασίας…….23

2.10.Διακύμανση μέγιστης Θερμοκρασίας…….24

2.11.Διακύμανση Βροχόπτωσης…….25

2.12. Διακύμανση μέγιστης ταχύτητας Ανέμου…….26

2.13. Διακύμανση κατεύθυνσης Ανέμου……27

2.14.Διακύμανση μέσης Υγρασίας…….28

2.15.Σύγκριση των τεσσάρων ασθενειών……...29

2.16. Ποσοστά χρονολογικής σύγκρισης του Κορωνοϊού……...30

2.17. Ποσοστά χρονολογικής σύγκρισης ανά Ηλικία…….31

2.18. Ποσοστά χρονολογικής σύγκρισης ανά Φύλο……...32

2.19. Ποσοστά χρονολογικής σύγκρισης ανά Επαρχία……...33

3.1.Κατάταξη της συσχέτισης ανά Φύλο……35

3.2. Κατάταξη της συσχέτισης ανά Ηλικία…….37

3.3. Κατάταξη της συσχέτισης ανά Επαρχία……...39

3.4. Καμπύλη ROC Curve……...42

3.5.Συνάρτηση επιβίωσης των ασθενών…….47

3.6.Συνάρτηση επιβίωσης των ασθενών με την μέθοδο Kaplan-Meier……..48

3.7. Συνάρτηση επιβίωσης των ασθενών ανά Φύλο……...48

3.8. Συνάρτηση επιβίωσης των ασθενών ανά Φύλο με την μέθοδο Kaplan-Meier…49

3.9. Συνάρτηση επιβίωσης των ασθενών ανά Ηλικία…….49

3.10. Συνάρτηση επιβίωσης των ασθενών ανά Ηλικία με την μέθοδο Kaplan-Meier……50

3.11. Συνάρτηση επιβίωσης των ασθενών ανά Επαρχία…….50

3.12. Συνάρτηση επιβίωσης των ασθενών ανά Επαρχία με την μέθοδο Kaplan-Meier……51

4.1. Διασπορά κρουσμάτων και μέσης Θερμοκρασίας……..56

4.2. Διασπορά κρουσμάτων και μέσης Υγρασίας……..57

4.3.Χρονοσειρά μέσης Θερμοκρασίας……58

4.4. Χρονοσειρά μέσης Υγρασίας……...58

4.5. Χρονοσειρά Κρουσμάτων……59

4.6.Συντελεστές και διαστήματα εμπιστοσύνης……..61

4.7.Συντελεστές και διαστήματα εμπιστοσύνης……..62

4.8. Συντελεστές και διαστήματα εμπιστοσύνης…….63

4.9. Συντελεστές και διαστήματα εμπιστοσύνης…….64

4.10.Μελλοντική πρόβλεψη Κρουσμάτων……..66

ΚΕΦΑΛΑΙΟ 1^0

Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε τον ορισμό του κορωνοϊού καθώς επίσης και κάποιες πληροφορίες σχετικά με την συμπεριφορά του ιού στον ανθρώπινο οργανισμό αλλά θα αναφερθούμε και στην μελέτη του Κέντρου Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDC).

1.1 Κορωνοϊος

1.1.1 Ιστορία του Κορωνοϊού

Οι κορωνοϊοί είναι ένα είδος ιών οι οποίοι προκαλούν αναπνευστικές λοιμώξεις στους ανθρώπους και στα ζώα. Οι επιστήμονες το 1968 ονόμασαν αυτούς τους ιούς ως κορωνοϊούς καθώς στο ηλεκτρονικό μικροσκόπιο η εικόνα τους έμοιαζε με κορώνα. Σύμφωνα με επιστημονικές μελέτες οι περισσότεροι άνθρωποι μολύνονται με κάποιο κορωνοϊό τουλάχιστον μια φορά στην ζωή τους, ωστόσο δεν έχει αποκαλυφθεί ακόμα ο λόγος για τον οποίο ο SARS-CoV-1, MERS-CoV και SARS-CoV-2 (γνωστός ως COVID-19) προκαλούν πιο σοβαρά συμπτώματα και έχουν μεγαλύτερο δείκτη θνησιμότητας σε σχέση με άλλους κορωνοϊούς.

Ο SARS-CoV-2 εμφανίστηκε για πρώτη φορά στην πόλη Wuhan(πρωτεύουσα της επαρχίας Hubei της Κίνας) στα τέλη Δεκεμβρίου 2019 και ως νέος ιός δεν υπήρχε ανοσία στον ανθρώπινο οργανισμό, με αποτέλεσμα ο Παγκόσμιος Οργανισμός Υγείας να κηρύξει τον COVID-19 ως πανδημία στις 11 Μαρτίου 2020 λόγω της ταχύτατης εξάπλωσής του.

1.1.2 Σύγκριση του Κορωνοϊού με τον ιό της γρίπης (H1N1)

Ο ιός της γρίπης και ο Covid-19 είναι δύο μεταδοτικές αναπνευστικές ασθένειες που προκαλούνται από διαφορετικούς ιούς, δηλαδή ο Covid-19 προκαλείται από τον νέο κορωνοϊό, τον λεγόμενο SARS-CoV-2 ενώ η γρίπη προκαλείται από ιούς της γρίπης. Οι δύο αυτοί ιοί παρουσιάζουν αρκετές ομοιότητες αλλά και διαφορές, συγκεκριμένα, ο Covid-19 φαίνεται να μεταδίδεται πιο γρήγορα και να προκαλεί πιο σοβαρά συμπτώματα στους ανθρώπους καθώς η διάρκεια μετάδοσης της ασθένειας είναι μεγαλύτερη σε σχέση με την γρίπη.

Μια σημαντική διαφορά είναι ότι έχει ανακαλυφθεί το εμβόλιο της γρίπης και έτσι περιορίζεται η μετάδοσή του ιού στον ανθρώπινο οργανισμό ενώ για τον Covid-19 μέχρι να ανακαλυφθεί το εμβόλιο ο καλύτερος τρόπος πρόληψης της λοίμωξης είναι η αποφυγή έκθεσης στον ιό. Ωστόσο και οι δύο ιοί παρουσιάζουν ομοιότητες ως προς τα συμπτώματα και τα χαρακτηριστικά με αποτέλεσμα να χρειάζονται δοκιμές για την εξακρίβωση της διάγνωσης μεταξύ των ασθενειών αυτών.

1.1.3 Συμπτώματα και μετάδοση στον ανθρώπινο οργανισμό

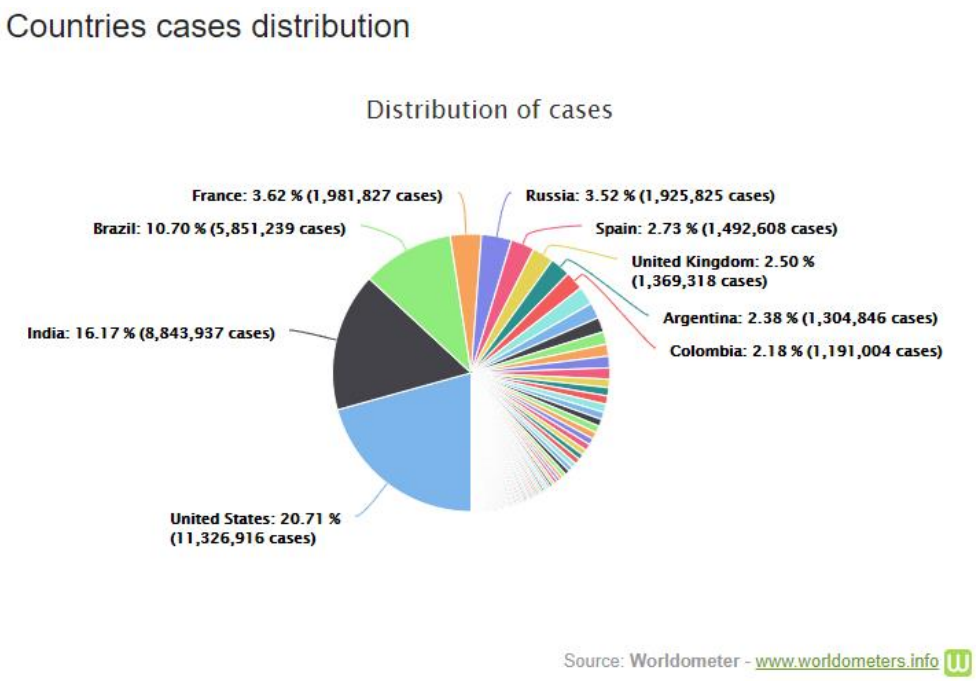
Ο κορωνοϊός είναι μια επικίνδυνη μεταδοτική ασθένεια η οποία μπορεί να έχει ήπια έως σοβαρά συμπτώματα στους ανθρώπους. Συγκεκριμένα τα ηλικιωμένα άτομα και τα άτομα που έχουν υποκείμενα νοσήματα όπως καρδιακή ή πνευματική νόσο ή διαβήτη διατρέχουν μεγαλύτερο κίνδυνο εμφάνισης σοβαρών επιπλοκών λόγω της ασθένειας Covid-19. Τα συμπτώματα μπορούν να εμφανιστούν 2-14 μέρες μετά την έκθεση στον ιό και ορισμένα από αυτά είναι ο πυρετός, πόνος στους μυς, πονοκέφαλος, βήχας και απώλεια γεύσης ή μυρωδιάς.

Ο ιός αυτός εξαπλώνεται από άτομο σε άτομο μέσω της στενής επαφής ακόμα και από άτομα που έχουν τον ιό στον οργανισμό τους αλλά δεν εμφανίζουν συμπτώματα, τα λεγόμενα ασυμπτωματικά άτομα. Η μετάδοση του ιού γίνεται με την μεταφορά αναπνευστικών σταγονιδίων από άτομο σε άτομο τα οποία παράγει όταν βήχει, φτερνίζεται, τραγουδάει , μιλάει ή αναπνέει. Είναι επίσης πιθανόν ένα άτομο να μολυνθεί με τον Covid-19 όταν αγγίξει επιφάνειες ή αντικείμενα που έχουν τον ιό πάνω τους και μετά αγγίξει το στόμα , την μύτη ή τα μάτια του. Ωστόσο, οι ενδείξεις για μετάδοση του ιού από ζώα σε ανθρώπους είναι χαμηλές ενώ υπάρχουν ορισμένες περιπτώσεις που ο άνθρωπος μπορεί να μεταδώσει τον ιό στα ζώα μέσω της στενής επαφής μαζί τους.

1.1.4 Στατιστικά στοιχεία και αντιμετώπιση

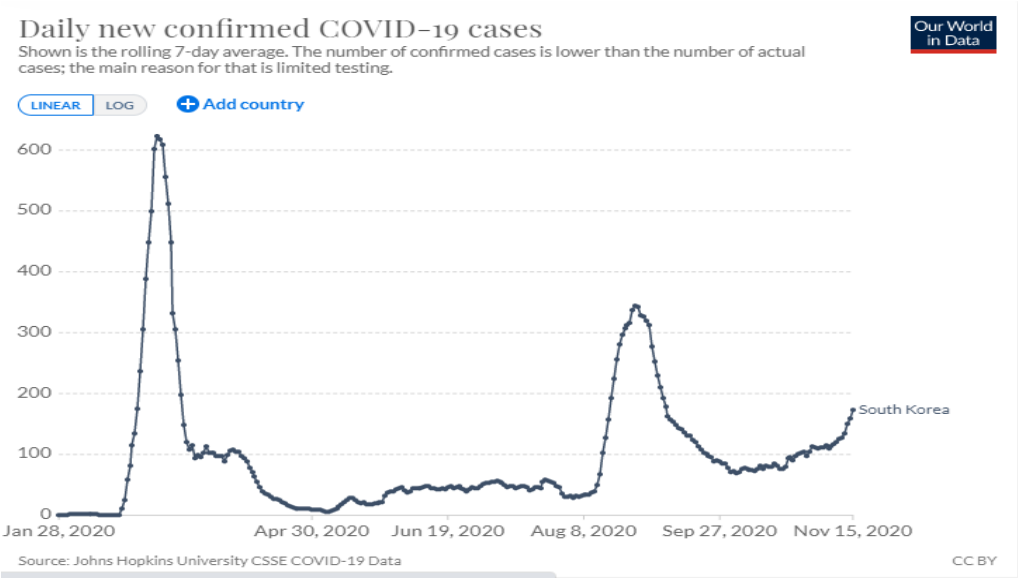
Για την πρόληψη της ασθένειας είναι σημαντικό να αποφεύγουμε την έκθεση στον ιό με το να φοράμε μάσκα ώστε να καλύπτετε το στόμα και η μύτη, να πλένουμε καλά τα χέρια μας, να κρατάμε αποστάσεις από τους γύρω μας, να απολυμάνουμε επιφάνειες τις οποίες έχουμε αγγίξει και να παρακολουθούμε συχνά την πορεία της υγείας μας. Εκτός από όλα αυτά πολλές είναι οι φαρμακευτικές οι οποίες συνέλαβαν στην δημιουργία εμβολίων (Pfizer, Moderna, AstraZeneca, Johnson & Johnson) τα οποία συμβάλλουν στην μείωση της εξάπλωσης της ασθένειας.

Είναι σημαντικό να τονίσουμε ότι τα στατιστικά στοιχεία δείχνουν ότι μέχρι 13 Νοεμβρίου 2020 , παγκοσμίως υπήρχαν 53.554.608 κρούσματα, 1.306.308 θανάτους και ότι είχαν αναρρώσει 37.415.987 εκατομμύρια άνθρωποι. Παρατηρούμε επίσης ότι οι χώρες με τον μεγαλύτερο ρυθμό μετάδοσης της ασθένειας ήταν η Αμερική , η Ινδία , η Βραζιλία, η Γαλλία και η Ρωσία.



1.1.5 Εμφάνιση του Κορωνοϊού στην Νότια Κορέα

Η Νότια Κορέα είναι η δεύτερη χώρα μετά την Κίνα στην οποία εμφανίστηκε ο Covid-19, με το πρώτο της κρούσμα να είναι στις 20 Ιανουαρίου 2020 το οποίο προερχόταν από την πόλη Wuhan της Κίνας. Ο αριθμός κρουσμάτων τον πρώτο μήνα έφταναν τα δύο ανά ημέρα, ωστόσο ένας ασθενής που ταξίδεψε στις πόλεις Daegu και Seoul πριν διαγνωστεί με την ασθένεια είχε σαν αποτέλεσμα να αυξηθούν τα κρούσματα ραγδαία και να φτάσουν 909 μέχρι τις 29 Φεβρουαρίου. Η γρήγορη αντιμετώπιση της κατάστασης από την κυβέρνηση μέσω των ειδικών τεστ ανίχνευσης του ιού στους πολίτες συνέβαλε στην μείωση των κρουσμάτων και στην σταθεροποίηση της κατάστασης έως ότου ανακαλυφθεί το εμβόλιο.



1.2 Μελέτη του Κέντρου Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDC)

1.2.1 Ιστορία του Οργανισμού Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA)

Το Κέντρο Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας το οποίο έχει ονομαστεί πλέον ως Οργανισμός Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA) είναι ένας οργανισμός που ιδρύθηκε στην Νότια Κορέα ο οποίος συμβάλλει στην πρόοδο της δημόσιας υγείας μέσω της πρόληψης και της έρευνας για μολυσματικές ασθένειες, τραυματισμούς, χρόνιες και σπάνιες παθήσεις. Ιδρύθηκε τον Δεκέμβριο του 2003 και συγκροτείται στην περιοχή Cheongju στο συγκρότημα διοίκησης τεχνολογίας υγείας Osong.

1.2.2 Τομείς δραστηριότητας του Οργανισμού Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA)

Ο συγκεκριμένος οργανισμός στοχεύει σε ένα ασφαλέστερο και υγιέστερο μέλλον για την κορεάτικη κοινωνία. Συγκεκριμένα, να διατηρήσει την κορεάτικη κοινωνία ασφαλής από απειλητικές ασθένειες ενισχύοντας τα μέτρα έκτακτης ανάγκης για την αντιμετώπιση μολυσματικών αυτών ασθενειών καθώς και στην εξέλιξη της έρευνας πάνω σε αυτές τις ασθένειες όπως και σε χρόνιες και σπάνιες ασθένειες. Επίσης, ειδικεύεται στην προώθηση της έρευνας και παρακολούθησης των χρόνιων παθήσεων καθώς επίσης και στην εξέλιξη της βιοϊατρικής επιστήμης στην Κορέα.

1.2.3 Αναφορά στην μελέτη του Οργανισμού Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA) για τον Κορωνοϊό.

Σύμφωνα με τον Οργανισμό Ελέγχου και Πρόληψης Νοσημάτων (KCDA) ο αριθμός των κρουσμάτων από τον κορωνοϊό είναι περισσότερα από 10.000 στην Νότια Κορέα μέχρι τις 20 Ιουνίου. Ο οργανισμός αυτός συγκεντρώνει τα δεδομένα της από τις τοπικές κυβερνήσεις καθώς και υλικό από τη βάση αναφοράς της, ανακοινώνοντας τις πληροφορίες γρήγορα και με διαφάνεια. Επίσης, χρησιμοποιεί τεχνικές της εξόρυξης δεδομένων ή οπτικοποίησης με τις οποίες αναλύει και οπτικοποιεί τα δεδομένα αυτά εξάγοντας σημαντικά συμπεράσματα για τη διασπορά του Covid-19 .

1.3 Δεδομένα

1.3.1 Στοιχεία της παρούσας εργασίας

Στην παρούσα ανάλυση τα αρχεία που θα χρησιμοποιήσουμε προέρχονται από τη βάση δεδομένων του Οργανισμού Ελέγχου και Πρόληψης Νοσημάτων της Νότιας Κορέας (KCDA) και αφορούν την περίοδο 20/01/2020 έως 29/06/2020.

Χρησιμοποιήθηκαν τα αρχεία

- Case
- Patient Info
- Policy
- Region
- Search Trend
- Time
- Time Age
- Time Gender
- Time Province
- Weather

Συγκεκριμένα, το αρχείο Case αναφέρεται στις περιπτώσεις οι οποίες μολύνθηκαν από τον Covid-19, δηλαδή για τον επιβεβαιωμένο αριθμό κρουσμάτων όπως επίσης στην επαρχία και την πόλη που εμφανίστηκε η μόλυνση συμπεριλαμβανομένου και της ακριβής τοποθεσίας των κρουσμάτων.

Το αρχείο PatientInfo αναφέρεται στα επιδημιολογικά στοιχεία για τους ασθενείς του κορωνοϊού, δηλαδή παρουσιάζει δημογραφικά στοιχεία των ασθενών όπως το φύλο και την ηλικία αλλά και σημαντικές πληροφορίες για την εμφάνιση και την εξέλιξη της ασθένειας σε κάθε άτομο.

Τα αρχεία Policy, Region, SearchTrend και Weather περιλαμβάνουν γενικές πληροφορίες για την Νότια Κορέα, δηλαδή το αρχείο Policy περιλαμβάνει στοιχεία σχετικά με την κυβερνητική πολιτική της χώρας. Τα στοιχεία από το αρχείο Region ειδικεύονται σε στατιστικά στοιχεία για τον πληθυσμό των ηλικιωμένων και την εκπαίδευση σε διάφορες πόλεις και επαρχίες της χώρας ενώ στο αρχείο SearchTrend περιλαμβάνονται στοιχεία σχετικά με τέσσερα είδη ασθένειας, δηλαδή το κρυολόγημα, την πνευμονία, την γρίπη και τον κορωνοϊό. Ένα ακόμη αρχείο είναι το Weather στο οποίο αναφέρονται περιβαλλοντικοί

παράγοντες όπως είναι η υγρασία και η θερμοκρασία στις διάφορες επαρχίες της Νότιας Κορέας.

Επιπρόσθετα, στην παρούσα εργασία υπάρχουν και αρχεία με χρονολογικά δεδομένα τα οποία είναι τα Time, TimeAge, TimeGender και TimeProvince.Το αρχείο Time περιλαμβάνει την κατάσταση στην οποία βρίσκεται χρονολογικά ο κορωνοϊός στην Νότια Κορέα ενώ το αρχείο TimeAge αναφέρεται στην πορεία της ασθένειας με βάση την ηλικία των ασθενών. Ωστόσο, και τα αρχεία TimeGender και TimeProvince περιλαμβάνουν χρονολογικά δεδομένα για τον Covid-19 αλλά το αρχείο TimeGender ειδικεύεται στην πορεία της εξάπλωσης με βάση το φύλο ενώ το αρχείο TimeProvince με βάση τις διάφορες επαρχίες της Νότιας Κορέας. Αναλυτικές πληροφορίες για τα αρχεία παρέχονται στο Παράρτημα 1 της παρούσας εργασίας.

1.4 Μεθοδολογία

1.4.1 Έλεγχος X^2

Η μέθοδος X^2 χρησιμοποιείται γενικά για να εκτιμήσουμε αν δύο ή περισσότερα δείγματα τα οποία αποτελούνται από δεδομένα συχνοτήτων διαφέρουν σημαντικά μεταξύ τους, δηλαδή χρησιμοποιείται κυρίως για την ανάλυση πινάκων διασταύρωσης ή συνάφειας με βάση δύο κατηγορικές μεταβλητές.

Γενικά, η μηδενική υπόθεση σε ένα πίνακα διασταύρωσης με r γραμμές και c στήλες είναι ότι δεν υπάρχει σχέση μεταξύ της μεταβλητής <<γραμμή>> και της μεταβλητής <<στήλη>>. Για να το ελέγξουμε αυτό συγκρίνουμε τις παρατηρούμενες συχνότητες (Π) κάθε κελιού με τις αναμενόμενες συχνότητες (Α) που υπολογίζονται υπό τη μηδενική υπόθεση. Πιο συγκεκριμένα, η αναμενόμενη τιμή κάθε κελιού εφόσον ισχύει η μηδενική υπόθεση δίνεται από τη σχέση

$$A = \frac{RT * CT}{N}$$

Όπου *RT* είναι το σύνολο της γραμμής που ανήκει το συγκεκριμένο κελί, *CT* το σύνολο της στήλης που ανήκει το συγκεκριμένο κελί και *N* ο συνολικός αριθμός των παρατηρήσεων.

Για να ελέγξουμε την μηδενική υπόθεση, υπολογίζουμε την τιμή της στατιστικής συνάρτησης χ^2 που συγκρίνει το σύνολο των παρατηρηθέντων συχνοτήτων με το σύνολο των αναμενόμενων συχνοτήτων, ως εξής:

$$X^2 = \sum \frac{(\Pi - A)^2}{\Pi}$$

με την άθροιση να γίνεται για όλα τα *r*c* κελία. Ο αριθμός των βαθμών ελευθερίας δίνεται από *df = (r-1) (c -1).*

Οι υποθέσεις που χρησιμοποιούνται είναι ως εξής:

H_0: Οι μεταβλητές *X* και *Y* είναι ανεξάρτητες

H_1: Οι μεταβλητές *X* και *Y* δεν είναι ανεξάρτητες

Ο κανόνας ορθής εφαρμογής του ελέγχου χ^2 έχει ως εξής:

- Το μέγεθος *n* του δείγματος δεν πρέπει να είναι μικρότερο του τετραπλάσιου του αριθμού των κελιών του πίνακα συνάφειας
- Καμία από τις αναμενόμενες συχνότητες (A) δεν πρέπει να είναι μικρότερη του 1
- Το ποσοστό των αναμενόμενων συχνοτήτων (A) οι οποίες είναι μικρότερες του 5 ή δεν πρέπει να είναι μεγαλύτερο του 20% με 25%.

Όταν σε ένα πίνακα συνάφειας ένα ή περισσότερα κελιά έχουν αναμενόμενη συχνότητα μικρότερη του 5, δηλαδή όταν παραβιάζεται ο τρίτος κανόνας ορθής εφαρμογής του ελέγχου χ^2 τότε αντί του ελέγχου χ^2 χρησιμοποιούμε τον ακριβή έλεγχο του Fisher.

1.4.2 Έλεγχος Kruskal-Wallis για κ ανεξάρτητα δείγματα

Ο έλεγχος *H* των Kruskal-Wallis χρησιμοποιείται αντί της Ανάλυσης Διακύμανσης με έναν παράγοντα, στην περίπτωση που οι υποθέσεις περί κανονικότητας των κ πληθυσμών δεν μπορεί να επαληθευτούν.

Ας υποθέσουμε ότι έχουμε λάβει δείγματα μεγέθους n_i , $1 \leq i \leq \kappa$, από κ ανεξάρτητους πληθυσμούς (θυμηθείτε ότι οι πληθυσμοί από τους οποίους προέρχονται τα δείγματα δηλώνονται ως επίπεδα μιας μεταβλητής κατηγορίας (παράγοντας)) και θέλουμε, χωρίς να υποθέσουμε ομοσκεδαστικότητα και κανονικότητα των πληθυσμών, να ελέγξουμε τις εξής υποθέσεις:

H_0: Οι *κ* πληθυσμοί έχουν την ίδια κατανομή

H_1: Οι *κ* πληθυσμοί δεν έχουν την ίδια κατανομή

Οι Kruskal-Wallis πρότειναν για αυτό τον έλεγχο μια μη παραμετρική διαδικασία βασισμένη στη βαθμολογία (Rank) των παρατηρήσεων.

Συγκεκριμένα, οι $N = n_1 + n_2 + \ldots + n_k$ συνολικές παρατηρήσεις που λάβαμε από τους κ πληθυσμούς ταξινομούνται από την μικρότερη στη μεγαλύτερη και στην κάθε μια δίνεται ένας βαθμός (Rank) ανάλογα με το μέγεθός της, ήτοι 1 στη μικρότερη, 2 στην αμέσως μεγαλύτερη κ.ο.κ. Σε περιπτώσεις ομάδων ίδιων παρατηρήσεων (ισοπαλίες-ties) η βαθμολογία αναπροσαρμόζεται δίνοντας σε κάθε παρατήρηση της ίδιας ομάδας το μέσο βαθμό που προκύπτει από τους αντίστοιχους αρχικούς βαθμούς. Στη συνέχεια υπολογίζουμε τις ποσότητες R_i , $1 \leq i \leq \kappa$, αθροίζοντας τις τελικές βαθμολογίες των παρατηρήσεων από κάθε δείγμα.

Η τιμή της στατιστικής συνάρτησης *H* των Kruskal-Wallis υπολογίζεται από τον τύπο

$$H = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)}{1 - \sum \frac{t^3 - t}{N^3 - N}}$$

όπου *N* το σύνολο των παρατηρήσεων. Το άθροισμα στον παρονομαστή γίνεται χάριν διόρθωσης ισοπαλιών και η άθροιση γίνεται για κάθε ομάδα ίσων παρατηρήσεων μεγέθους *t*. Αν δεν υπάρχουν ίσες παρατηρήσεις, τότε κάθε μια θα αποτελεί μια ομάδα μεγέθους *t =1* και ο παρανομαστής που προκύπτει ισούται με τη μονάδα, δηλαδή προκύπτει η πιο απλή μορφή της συνάρτησης του ελέγχου Kruskal-Wallis.

1.4.3 Πολλαπλή Λογιστική Παλινδρόμηση

Στην περίπτωση της λογιστικής παλινδρόμησης η εξαρτημένη μεταβλητή *Y* είναι δίτιμη μεταβλητή, δηλαδή παίρνει δυο τιμές που αντιπροσωπεύουν συνήθως την παρουσία ή απουσία ενός χαρακτηριστικού. Σε τέτοιες περιπτώσεις σκοπός μας είναι να καθορίσουμε ορισμένες ανεξάρτητες μεταβλητές, οι οποίες απαιτούνται για την πρόβλεψη της μέσης τιμής της δίτιμης εξαρτημένης μεταβλητής. Αυτό επιτυγχάνεται μέσω της μεθόδου της λογιστικής παλινδρόμησης που χρησιμοποιείται για να περιγράψει τη σχέση της πιθανότητας ενός χαρακτηριστικού με διάφορους παράγοντες.

Έστω ότι έχουμε δυο ή περισσότερες ανεξάρτητες μεταβλητές και θέλουμε να ελέγξουμε εάν αυτές επηρεάζουν μια δίτιμη ερμηνευτική μεταβλητή. Το μοντέλο της λογιστικής παλινδρόμησης στην περίπτωση αυτή είναι ως εξής:

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_pX_{pi} + \varepsilon$$

και ονομάζεται μοντέλο πολλαπλής λογιστικής παλινδρόμησης.

Η ερμηνεία των παραμέτρων του παραπάνω μοντέλου είναι παρόμοια με αυτή των παραμέτρων του μοντέλου πολλαπλής γραμμικής παλινδρόμησης και για να αξιολογήσουμε τη στατιστική σημαντικότητα των παραμέτρων του μοντέλου της πολλαπλής λογιστικής παλινδρόμησης, συγκρίνουμε την τιμή της κάθε παραμέτρου με το τυπικό της σφάλμα. Το κριτήριο είναι γνωστό ως κριτήριο του Wald και οι υποθέσεις που ελέγχονται είναι οι εξής

$H_0: b_i = 0$

$H_1: b_i \neq 0$

Η μηδενική υπόθεση αναφέρει ότι η *i* ανεξάρτητη μεταβλητή δεν ερμηνεύει τον λογάριθμο του λόγου συμπληρωματικών πιθανοτήτων, log(OR) , ενώ η

εναλλακτική υπόθεση αναφέρει ότι η *i* ανεξάρτητη μεταβλητή ερμηνεύει τον λογάριθμο του λόγου συμπληρωματικών πιθανοτήτων.

Η στατιστική συνάρτηση του ελέγχου είναι

$$W = \frac{\hat{b}_i}{s(\hat{b}_i)}$$

η οποία, όταν ισχύει η μηδενική υπόθεση ακολουθεί την τυπική κανονική κατανομή *N(0,1)*.

1.4.4 Καμπύλες ROC

Η διάγνωση μιας ασθένειας γίνεται ορισμένες φορές με βάση μια συνεχή μέτρηση, όμως δεν υπάρχει κάποια διαχωριστική τιμή (cut-off point) πάνω ή κάτω από την οποία εμφανίζεται η ασθένεια. Για αυτό το λόγο για να προκύψει η συγκεκριμένη διαχωριστική τιμή θα πρέπει να ελαχιστοποιηθεί ο αριθμός των ψευδών θετικών και ψευδών αρνητικών αποτελεσμάτων χρησιμοποιώντας την ανάλυση ROC. Η ελαχιστοποίηση των ψευδώς θετικών και ψευδώς αρνητικών αποτελεσμάτων είναι το ίδιο με τη μεγιστοποίηση της ευαισθησίας και της ειδικότητας.

Η καμπύλη ROC (Receiver Operating Characteristic curve) είναι μια γραφική παράσταση της ευαισθησίας έναντι της ποσότητας (1-ειδικότητα). Στον κατακόρυφο άξονα αναπαριστάται η ευαισθησία, δηλαδή η πιθανότητα ο έλεγχος να δώσει θετικό αποτέλεσμα δοθέντος ότι το άτομο πάσχει από την ασθένεια, ενώ στον οριζόντιο άξονα αναπαριστάται η ποσότητα (1-ειδικότητα), δηλαδή η πιθανότητα ο έλεγχος να δώσει θετικό αποτέλεσμα δοθέντος ότι το άτομο δεν πάσχει από την ασθένεια.

Η μεγιστοποίηση της ευαισθησίας αντιστοιχεί σε μια μεγάλη τιμή στον κατακόρυφο άξονα ενώ η μεγιστοποίηση της ειδικότητας αντιστοιχεί σε μια μικρή τιμή στον οριζόντιο άξονα. Έτσι μια καλή πρώτη επιλογή είναι για την τιμή της αποκοπής του διαγνωστικού ελέγχου είναι μια τιμή που αντιστοιχεί σε ένα σημείο της καμπύλης ROC που βρίσκεται κοντά στην επάνω αριστερή γωνία του διαγράμματος ROC.

Για την εύρεση του βέλτιστου σημείου αποκοπής *c* μεγιστοποιούμε τη συνάρτηση

$$h(c) = \text{ευαισθησία} + \text{ειδικότητα} - 1,$$

η οποία είναι γνωστή ως Δείκτης του Youden.

Για τη σύγκριση διαφορετικών διαγνωστικών ελέγχων και για την επιλογή τιμών αποκοπής, χρησιμοποιείται το εμβαδόν κάτω από την καμπύλη (area under curve-AUC). Είναι προφανές ότι όσο η καμπύλη πλησιάζει την επάνω αριστερή γωνία του τετραγώνου της γραφικής παράστασης, δηλαδή όσο μεγαλύτερο είναι το εμβαδόν κάτω από την καμπύλη (AUC) τόσο πιο αξιόπιστος είναι ο διαγνωστικός

έλεγχος. Συνήθως, το σημείο που είναι το πλησιέστερο σε αυτή τη γωνία είναι αυτό που επιλέγεται ως σημείο αποκοπής καθώς αυτό μεγιστοποιεί ταυτόχρονα και την ευαισθησία και την ειδικότητα του ελέγχου.

1.4.5 Πίνακες Επιβίωσης με τη μέθοδο Kaplan-Meier

Όταν το δείγμα μας περιέχει πλήρεις και λογοκριμένους χρόνους ζωής η εμπειρική συνάρτηση επιβίωσης δεν μπορεί να χρησιμοποιηθεί ως εκτίμηση της *S(t)* επειδή δεν είναι γνωστός ο πραγματικός αριθμός των χρόνων ζωής που είναι μεγαλύτεροι ή ίσοι του *t* λόγω των διαφυγών που τυχόν έχουν παρατηρηθεί στο διάστημα *[0,t)*. Οι Kaplan & Meier πρότειναν τον εκτιμητή Kaplan-Meier ή εκτιμητή γινομένου ορίου (product limit estimator, PLE) ως εκτιμητής της *S(t)*. Ο PLE της *S(t)* ορίζεται ως ακολούθως:

Ας υποθέσουμε ότι υπάρχουν k $(k \leq n)$ διαφορετικές χρονικές στιγμές $t_1<t_2<\ldots<t_k$ στις οποίες συμβαίνουν οι θάνατοι (αποτυχίες), που αντιστοιχούν σε πλήρεις χρόνους ζωής και για $j=1,2,\ldots,k$ έστω

- d_j ο αριθμός των θανάτων που συμβαίνουν τη χρονική στιγμή t_j
- c_j ο αριθμός των διαφυγών στο διάστημα $[t_j,t_{j+1})$ και ορίζουμε αξιωματικά ότι $t_{k+1}=\infty$
- r_j ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο τη χρονική στιγμή t_j

Ο PLE που πρότειναν οι Kaplan-Meier δίνεται από την σχέση

$$\hat{s}(t) = \prod_{j:t_j<t} \frac{r_j - d_j}{d_j} = \prod_{j:t_j<t} \left(1 - \frac{d_j}{r_j}\right)$$

Όπου ως συνήθως $\hat{s}(t) = 1$ για $t \leq t_1$.

1.4.6 Το μοντέλο αναλογικού κινδύνου (PH) του Cox

Τα δεδομένα που έχουμε στη διάθεσή μας για τα n υπό παρακολούθηση άτομα εκτός από την τιμή της τυχαίας μεταβλητής T (που δηλώνει χρόνο ζωής) και την τιμή της τυχαίας μεταβλητής Δ (που δηλώνει αν ο χρόνος ζωής είναι πλήρης ή λογοκριμένος) συνοδεύονται και από τιμές $Z=(Z_1,Z_2,\ldots,Z_p)'$ άλλων ερμηνευτικών μεταβλητών (explanatory variables), ή συμμεταβλητών (covariates), ή απλά μεταβλητών, που δίνουν πληροφορίες για διάφορα άλλα χαρακτηριστικά των ατόμων. Στην περίπτωση αυτή συγκρίνουμε συναρτήσεις επιβίωσης διαφόρων ομάδων λαμβάνοντας υπόψη τις τιμές δυο ή περισσοτέρων συμμεταβλητών για να διαπιστώσουμε τη σχέση μεταξύ της μεταβλητής που δηλώνει το χρόνο ζωής και μιας ή περισσοτέρων συμμεταβλητών που επιτυγχάνεται μέσω ενός μοντέλου παλινδρόμησης.

Στόχος του είναι ο καθορισμός των συμμεταβλητών που επηρεάζουν τη συνάρτηση κινδύνου και η εκτίμηση της συνάρτησης κινδύνου και συνεπώς η εκτίμηση της συνάρτησης επιβίωσης. Το βασικό μοντέλο παλινδρόμησης που χρησιμοποιείται στη ανάλυση επιβίωσης είναι το μοντέλο αναλογικού κινδύνου ή μοντέλο PH, το οποίο αναφέρεται και ως μοντέλο παλινδρόμησης του Cox. Είναι ένα ημι-παραμετρικό μοντέλο αφού μόνο το διάνυσμα των συμμεταβλητών *Z* εισέρχεται στο μοντέλο με παραμετρική μορφή ενώ για την αναφορική συνάρτηση κινδύνου δε γίνεται καμία παραμετρική υπόθεση.

Ο Cox πρότεινε την μορφή $C(x)= e^x$ για τη συνάρτηση $c(\cdot)$, οπότε

$$C(\beta'Z) = \exp(\beta'Z) = \exp\left(\sum_{k=1}^{p}\beta_k z_k\right)$$

και

$$h(t|Z) = h_0(t)\exp(\beta'Z) = h_0(t)\exp\left(\sum_{k=1}^{p}\beta_k z_k\right)$$

που είναι η πλήρης μορφή του αποκαλούμενου μοντέλου PH.

1.4.7 Συντελεστής συσχέτισης του Pearson

Έστω ότι δύο μεταβλητές *X* και *Y* για τις οποίες έχουμε καταγράψει τις τιμές χ_i της *X* και τις τιμές y_i της *Y* για $i= 1,2,\ldots,v$. Από την περιγραφική στατιστική είναι γνωστό ότι ο συντελεστής συσχέτισης των δύο μεταβλητών περιγράφεται από την σχέση

$$r_{x,y} = \frac{\sum_{i=1}^{v}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{v}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{v}(y_i-\bar{y})^2}}$$

και εκφράζει το βαθμό (γραμμικής) συσχέτισης μεταξύ των μεταβλητών *X* και *Y*, παίρνοντας τιμές μεταξύ του -1 και του 1. Όσο μεγαλύτερος είναι ο συντελεστής συσχέτισης $r_{x,y}$ τόσο μεγαλύτερη και η γραμμική εξάρτηση των δύο μεταβλητών.

1.4.8 Ανάλυση Χρονολογικών Σειρών

Χρονολογική σειρά είναι μια σειρά επαναλαμβανόμενων παρατηρήσεων μιας μεταβλητής, οι οποίες παρατηρούνται σε τακτικά διαστήματα κατά τη διάρκεια ενός συγκεκριμένου χρόνου.

Η χρονολογική σειρά έχεις τις εξής ιδιότητες:

- Τάση (trend). Είναι μια ανοδική ή καθοδική πορεία της σειράς κατά τη διάρκεια μακρού χρόνου. Η σειρά χωρίς τάση ονομάζεται στάσιμη σειρά (stationary series).
- Κυκλική διαφόριση (cyclical variation). Είναι μια τακτική πορεία ανοδικών ή καθοδικών κυμάτων κατά τη διάρκεια μιας μακράς περιόδου χρόνων.

- Εποχιακή διαφόριση (seasonal variation). Είναι μια πορεία ανόδου και καθόδου εντός ενός έτους.
- Ακανόνιστη διαφόριση (random ή irregular variation). Είναι ένα σύνολο μη κανονικών κινήσεων των παρατηρήσεων.

1.4.9 Η μεθοδολογία Box και Jenkins

Η μέθοδος Box και Jenkins χρησιμοποιείται για την εύρεση και εκτίμηση της καλύτερης ARIMA (p,d,q) διαδικασίας με βάση το δείγμα της χρονοσειράς που έχουμε διαθέσιμο. Οι Box και Jenkins προτείνουν τρία στάδια για την ανάλυση μιας χρονοσειράς, δηλαδή την ταυτοποίηση της χρονοσειράς ως ARIMA (p,d,q) διαδικασία, την εκτίμηση των παραμέτρων *p* και *q* του υποδείγματος καθώς επίσης και τον έλεγχο του συγκεκριμένου υποδείγματος.

1) Ταυτοποίηση της χρονοσειράς

Αρχικά θα πρέπει να ελέγξουμε αν οι παρατηρήσεις προέρχονται από στάσιμη σειρά το οποίο μπορούμε να ελέγξουμε διαγραμματικά αλλά και χρησιμοποιώντας τις αυτοσυσχετίσεις και μερικές αυτοσυσχετίσεις του υποδείγματός μας καθώς και τον έλεγχο Box-Ljung που αναφέρεται στην ύπαρξη ή μη συσχέτισης των τιμών της χρονοσειράς, τον λεγόμενο λευκό θόρυβο.

2) Εκτίμηση των παραμέτρων p και q του υποδείγματος

Από το προηγούμενο στάδιο αποφασίζουμε για τις τιμές των *p* και *q* και στην συνέχεια εκτιμούμε τους συντελεστές του υποδείγματος. Η εκτίμηση του υποδείγματος γίνεται είτε μέσω της μεθόδου της μέγιστης πιθανοφάνειας είτε της μεθόδου των ελαχίστων τετραγώνων χωρίς συνθήκη είτε της μεθόδου των ελαχίστων τετραγώνων με συνθήκη. Εφαρμόζοντας μία από τις τρεις συνθήκες μπορούμε να εκτιμήσουμε τις παραμέτρους, τις διακυμάνσεις τους καθώς επίσης να υπολογίσουμε διαστήματα εμπιστοσύνης και να διεξαχθούν έλεγχοι στατιστικής σημαντικότητας.

3)Έλεγχος του υποδείγματος

Στο τελευταίο στάδιο ασχολούμαστε με τον έλεγχο της επάρκειας του εκτιμηθέντος υποδείγματος το οποίο γίνεται μέσω των σφαλμάτων του υποδείγματος. Συγκεκριμένα θα πρέπει τα σφάλματα να αποτελούν λευκό θόρυβο αν το υπόδειγμά μας είναι ικανοποιητικό, το οποίο ελέγχουμε διαγραμματικά μέσω των αυτοσυσχετίσεων των εκτιμηθέντων σφαλμάτων αλλά και μέσω της στατιστικής συνάρτησης *Q* των Box και Ljung. Η στατιστική συνάρτηση *Q* των Box και Ljung ελέγχει την ταυτόχρονη στατιστική σημαντικότητα ενός συνόλου αυτοσυσχετίσεων των σφαλμάτων η οποία υπολογίζεται από την σχέση

$$Q_m = n\,(n+2) \sum_{i=1}^{m} \frac{r_i^2}{n-i}$$

όπου

H_0: $\rho_1=\rho_2=\ldots=\rho_m=0$

H_1: Μια τουλάχιστον από τις $\rho_i \neq 0$, για $i=1,2,\ldots,m$

ΚΕΦΑΛΑΙΟ 2^0

2.1 Περιγραφική Ανάλυση

Στο κεφάλαιο αυτό θα ασχοληθούμε εκτενέστερα με την περιγραφική ανάλυση των δημογραφικών και περιβαλλοντικών παραγόντων που συμβάλλουν στην εξάπλωση του κορωνοϊού στην Νότια Κορέα τους οποίους αναφέραμε επιγραμματικά στο προηγούμενο κεφάλαιο.

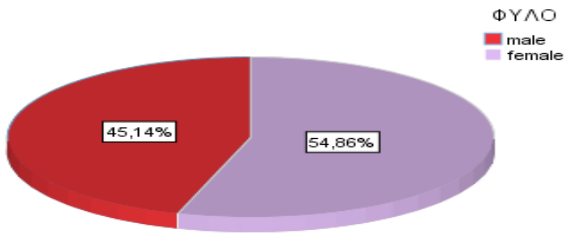
2.1.1 Δημογραφικοί Παράγοντες

2.1.1.1 Φύλο

ΦΥΛΟ					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	1825	35,3	45,1	45,1
	female	2218	42,9	54,9	100,0
	Total	4043	78,3	100,0	
Missing	System	1122	21,7		
Total		5165	100,0		

Πίνακας 2.1 Περιγραφή του Φύλου

Σύμφωνα με τον πίνακα έχουμε 1825 άνδρες ασθενείς και 2218 γυναίκες ασθενείς των οποίων τα ποσοστά φαίνονται στον παρακάτω διάγραμμα αλλά υπάρχουν 1122 ασθενείς για τους οποίους δεν έχει καταγραφεί το φύλο τους και θεωρούνται ελλιπείς τιμές.



Διάγραμμα 2.1. Ποσοστά ασθενών ανά Φύλο

2.1.1.2 **Ηλικία**

Για την ανάλυσή μας η ηλικία των ασθενών έχει χωριστεί στις εξής ομάδες:

- 0s: Ηλικίες από 0 έως 9
- 10s: Ηλικίες από 10 έως 19
- 20s: Ηλικίες από 20 έως 29
- 30s: Ηλικίες από 30 έως 39
- 40s: Ηλικίες από 40 έως 49
- 50s: Ηλικίες από 50 έως 59
- 60s: Ηλικίες από 60 έως 69
- 70s: Ηλικίες από 70 έως 79
- 80s: Ηλικίες από 80 έως 89
- 90s: Ηλικίες από 90 έως 99
- 100s: Ηλικίες από 100 έως 109

ΗΛΙΚΙΑ					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0s	66	1,3	1,7	1,7
	10s	178	3,4	4,7	6,4
	20s	899	17,4	23,8	30,2
	30s	523	10,1	13,8	44,0
	40s	518	10,0	13,7	57,7
	50s	667	12,9	17,6	75,3
	60s	482	9,3	12,7	88,1
	70s	232	4,5	6,1	94,2
	80s	170	3,3	4,5	98,7
	90s	49	,9	1,3	100,0
	100s	1	,0	,0	100,0
	Total	3785	73,3	100,0	
Missing	System	1380	26,7		
Total		5165	100,0		

Πίνακας 2.2.Περιγραφή των κατηγοριών Ηλικίας

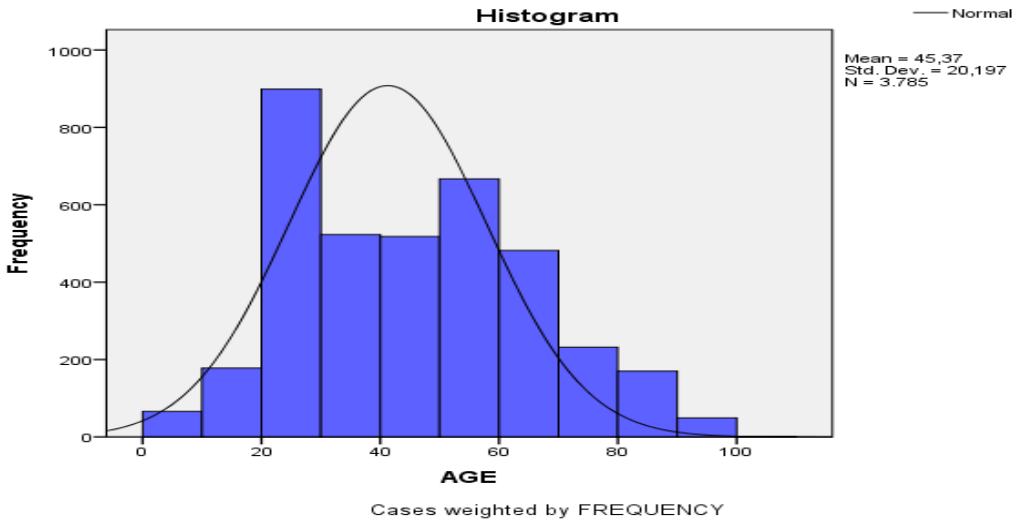
Παρατηρούμε ότι το μεγαλύτερο ποσοστό ασθενών το οποίο είναι 17,4% βρίσκεται στις ηλικίες 20 έως 29 ενώ το μικρότερο ποσοστό που είναι και μηδενικό βρίσκεται στις ηλικίες 100 έως 109. Από τον παραπάνω πίνακα

παρατηρούμε μεγαλύτερα ποσοστά στις ηλικίες 20 έως 59 ετών ενώ όσο προχωράμε σε μεγαλύτερη ηλικιακή ομάδα ασθενών τα ποσοστά σταδιακά ελαττώνονται.

Statistics		
AGE		
N	Valid	3785
	Missing	0
Mean		45,37
Median		45,00
Std. Deviation		20,197
Variance		407,925
Skewness		,307
Std. Error of Skewness		,040
Kurtosis		-,659
Std. Error of Kurtosis		,080
Minimum		5
Maximum		105

Πίνακας 2.3.Περιγραφικά μέτρα για την Ηλικία

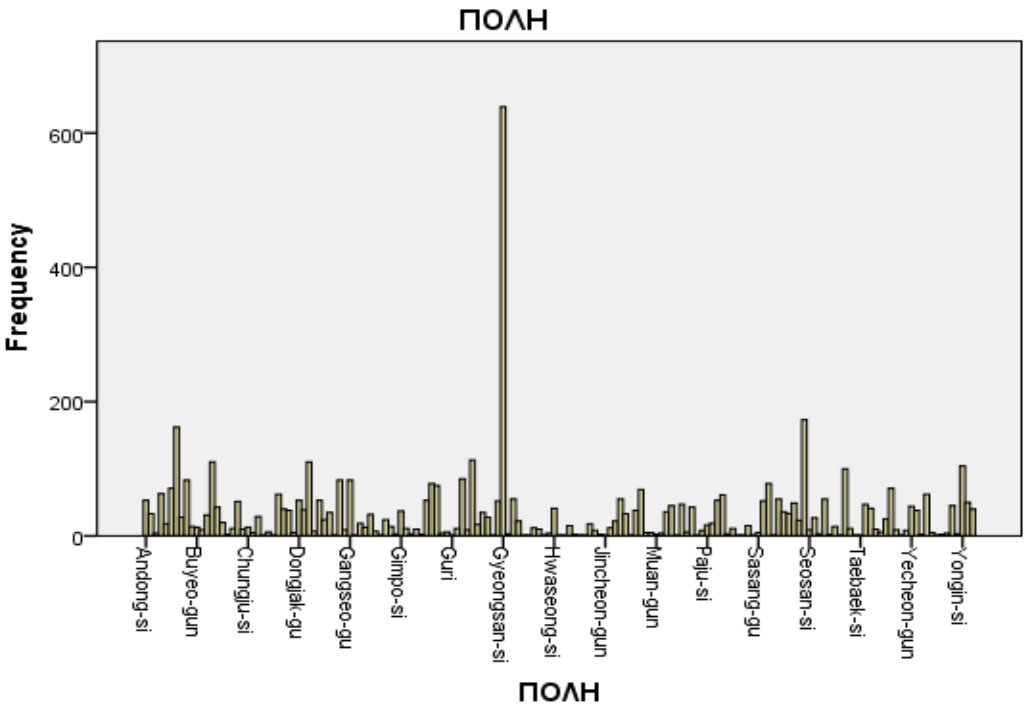
Το παραπάνω συμπέρασμα παρατηρείται και στον παρακάτω ιστόγραμμα διότι ο συντελεστής ασυμμετρίας (skewness=0,307) έχει θετικό πρόσημο δηλαδή έχουμε ελαφρώς θετική ασυμμετρία που σημαίνει ότι η εξάπλωση της ασθένειας κυμαίνεται γύρω από τις μικρότερες ηλικιακές ομάδες ασθενών όπως επίσης ότι η κατανομή μας είναι πλατύκυρτη (kurtosis=-0,659<3) που σημαίνει ότι η κυρτότητα των ηλικιών είναι μεγαλύτερη σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής και άρα συγκέντρωση τιμών γύρω από το μέσο.



Διάγραμμα 2.2. Διακύμανση της ασθένειας με βάση την Ηλικία

2.1.1.3 <u>Τόπος Κατοικίας</u>

Σύμφωνα με τον πίνακα συχνοτήτων παρατηρούμε ότι υπάρχουν μικρά ποσοστά όσον αφορά την πόλη στην οποία κατοικεί άτομο που νοσεί από τον κορωνοϊό όπως είναι η Gwanak-gu με ποσοστό 2,2% και η Seongnam-si με ποσοστό 3,3% αλλά ένα εμφανές μεγαλύτερο ποσοστό των ασθενών που νοσούν παρατηρείται ότι κατοικούν στην περιοχή Gyeongsan-si σε ποσοστό 12,4% το οποίο φαίνεται και έντονα στο παρακάτω ραβδόγραμμα. Ο πίνακας συχνοτήτων περιλαμβάνεται αναλυτικά στο Παράρτημα 2.



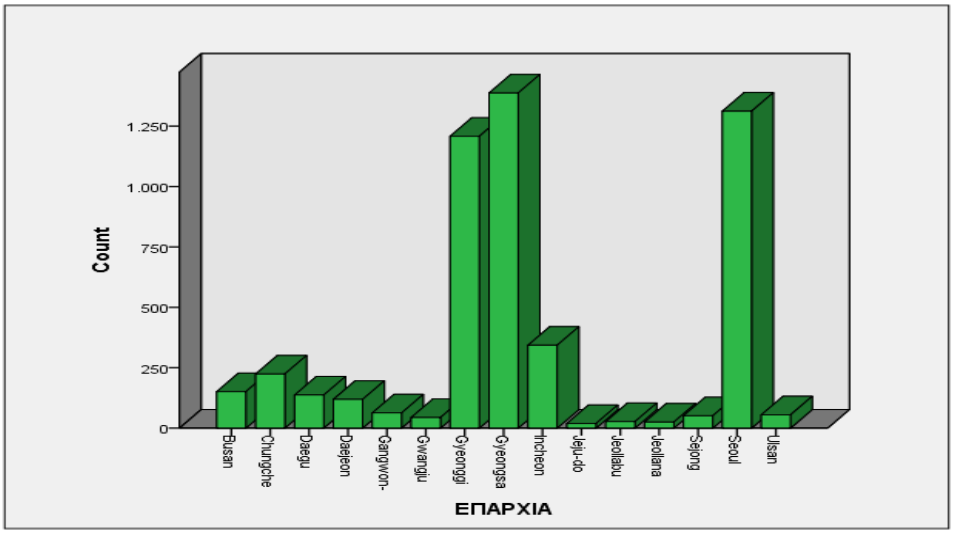
Διάγραμμα 2.3. Κατάταξη εμφάνισης της ασθένειας ανά Πόλη

2.1.1.4 Επαρχία Προέλευσης

ΕΠΑΡΧΙΑ					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Busan	151	2,9	2,9	2,9
	Chungche	224	4,3	4,3	7,3
	Daegu	137	2,7	2,7	9,9
	Daejeon	119	2,3	2,3	12,2
	Gangwon-	63	1,2	1,2	13,4
	Gwangju	44	,9	,9	14,3
	Gyeonggi	1208	23,4	23,4	37,7
	Gyeongsa	1387	26,9	26,9	64,5
	Incheon	343	6,6	6,6	71,2
	Jeju-do	19	,4	,4	71,5
	Jeollabu	27	,5	,5	72,1
	Jeollana	25	,5	,5	72,5
	Sejong	51	1,0	1,0	73,5
	Seoul	1312	25,4	25,4	98,9
	Ulsan	55	1,1	1,1	100,0
	Total	5165	100,0	100,0	

Πίνακας 2.4.Περιγραφή των διάφορων Επαρχιών

Σύμφωνα με τον πίνακα συχνοτήτων οι περισσότεροι ασθενείς συγκεντρώνονται σε τρεις επαρχίες, δηλαδή στην Gyeongsa με ποσοστό 26,9%, στην Seoul με ποσοστό 25,4% και στην επαρχία Gyeonggi με ποσοστό 23,4%, αντιθέτως στις υπόλοιπες επαρχίες κατοικούν ελάχιστοι ασθενείς.



Διάγραμμα 2.4.Κατάταξη των ασθενών ανά Επαρχία

Ωστόσο εξετάζοντας χρονικά την αλλαγή στον αριθμό κρουσμάτων στο σύνολο των επαρχιών ανά εβδομάδα χρησιμοποιώντας τον τύπο:

$$X = \frac{Μεταγενέστερη\ χρονική\ στιγμή - Προγενέστερη\ χρονική\ στιγμή}{Προγενέστερη\ χρονική\ στιγμή}$$

, καταλήγουμε στον παρακάτω πίνακα.

ΕΒΔΟΜΑΔΕΣ	ΚΡΟΥΣΜΑΤΑ ΑΝΑ ΕΒΔΟΜΑΔΕΣ	ΡΥΘΜΟΣ ΜΕΤΑΒΟΛΗΣ
1	11	11
2	16	0,454545
3	3	-0,812500
4	263	86,666667
5	756	1,874525
6	652	-0,137566
7	355	-0,455521
8	347	-0,022535
9	347	0,000000
10	349	0,005764
11	149	-0,573066
12	87	-0,416107
13	50	-0,425287
14	37	-0,260000
15	92	1,486486
16	129	0,402174
17	126	-0,023256
18	277	1,198413
19	247	-0,108303
20	282	0,141700
21	268	-0,049645
22	269	0,003731

Πίνακας 2.5.Ρυθμός μεταβολής Κρουσμάτων

Διάγραμμα 2.5.Ρυθμός μεταβολής κρουσμάτων ανά εβδομάδα

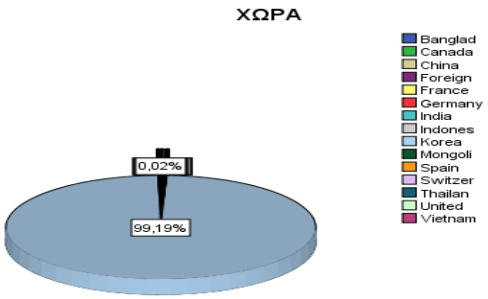
Σύμφωνα με το διάγραμμα χρόνου παρατηρείται ότι έντονη αύξηση των κρουσμάτων υπήρχε μεταξύ της 3ης και 5ης εβδομάδας ενώ με την πάροδο του χρόνου υπήρχε σταδιακή μείωση της έξαρσης της ασθένειας.

2.1.1.5 Χώρα Προέλευσης

ΧΩΡΑ					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Banglad	5	,1	,1	,1
	Canada	1	,0	,0	,1
	China	11	,2	,2	,3
	Foreign	7	,1	,1	,5
	France	1	,0	,0	,5
	Germany	1	,0	,0	,5
	India	1	,0	,0	,5
	Indones	2	,0	,0	,6
	Korea	5123	99,2	99,2	99,7
	Mongoli	1	,0	,0	99,8
	Spain	1	,0	,0	99,8
	Switzer	1	,0	,0	99,8
	Thailan	2	,0	,0	99,8
	United	7	,1	,1	100,0
	Vietnam	1	,0	,0	100,0
	Total	5165	100,0	100,0	

Πίνακας 2.6.Περιγραφή των διάφορων Χωρών

Στην ανάλυση συμπεραίνουμε ότι σχεδόν όλοι οι ασθενείς προέρχονται από την Κορέα με ποσοστό 99,2% και ελάχιστοι από αυτούς προέρχονται είτε από την Κίνα, είτε από το Μπανγκλαντές και είτε από άλλες ξένες χώρες το οποίο παρατηρείται έντονα και στο ραβδόγραμμα.



Διάγραμμα 2.6.Ποσοστά ασθενών ανά Χώρα

2.1.1.6 Μέρος εμφάνισης του κορωνοϊού

Λόγω ότι υπάρχουν στα δεδομένα μας αρκετά μέρη όπου εμφανίστηκε ο COVID-19, ο πίνακας συχνοτήτων χωρίζετε σε τέσσερις κατηγορίες, όπου:

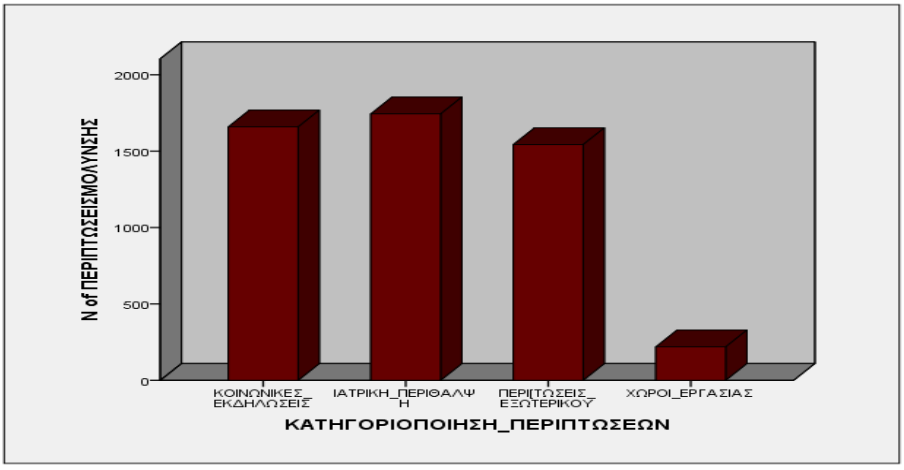
- 1η κατηγορία: Κοινωνικές εκδηλώσεις
- 2η κατηγορία: Ιατρική Περίθαλψη
- 3η κατηγορία: Περιπτώσεις Εξωτερικού
- 4η κατηγορία: Χώροι Εργασίας

Σύμφωνα με τον πίνακα συχνοτήτων κρίνονται σημαντικά τρεις καταστάσεις σύμφωνα με τις οποίες εμφανίζεται η εξάπλωση του COVID-19 στην Νότια Κορέα. Συγκεκριμένα, η ανάλυση αναφέρεται στο ότι 32,1% των ασθενών ασθένησαν λόγω της επαφής με άλλου ασθενείς, 16,3% των ασθενών μολύνθηκαν λόγω της επαφής με άτομα που προέρχονται από άλλες χώρες και το 13,6% αναφέρεται σε μεμονωμένες περιπτώσεις, σε περιπτώσεις που είναι υπό εξέταση καθώς επίσης και σε περιπτώσεις που συνδέονται με την συγκεκριμένη ασθένεια αφού πρώτα εξεταστούν. Ο πίνακας συχνοτήτων περιλαμβάνεται αναλυτικά στο παράρτημα δύο.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ_ΠΕΡΙΠΤΩΣΕΩΝ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΚΟΙΝΩΝΙΚΕΣ_ΕΚΔΗΛΩΣΕΙΣ	1659	32,1	32,1	32,1
	ΙΑΤΡΙΚΗ_ΠΕΡΙΘΑΛΨΗ	1744	33,8	33,8	65,9
	ΕΞΩΤΕΡΙΚΑ_ΚΡΟΥΣΜΑΤΑ	1543	29,9	29,9	95,8
	ΧΩΡΟΙ ΕΡΓΑΣΙΑΣ	219	4,2	4,2	100,0
	Total	5165	100,0	100,0	

Πίνακας 2.7.Περιγραφή σημείων εμφάνισης της ασθένειας



Διάγραμμα 2.7.Κατάταξη σημείων εμφάνισης της ασθένειας

Σύμφωνα με το ραβδόγραμμα παρατηρούμε ότι ο μεγαλύτερος αριθμός κρουσμάτων εμφανίζεται στην κατηγορία της ιατρικής περίθαλψης όπου περιέχει νοσοκομεία και ιατρικά κέντρα ενώ ελάχιστος αριθμός περιπτώσεων εμφανίζεται σε χώρους εργασίας.

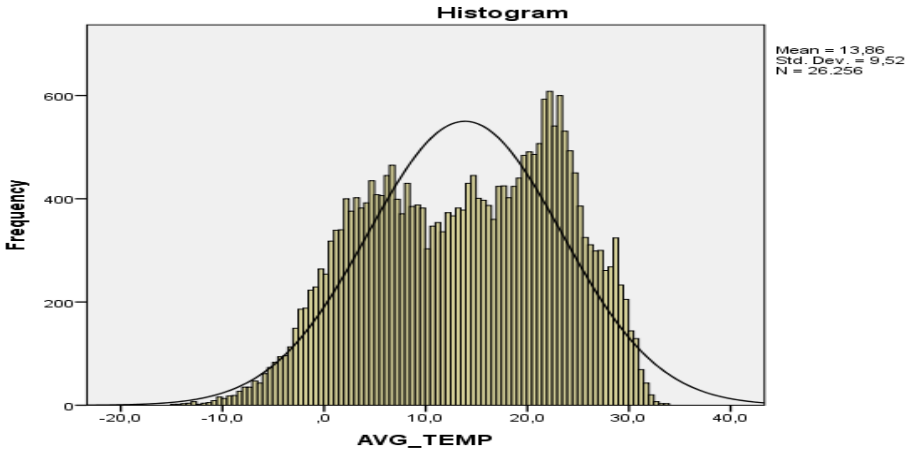
2.1.2 Περιβαλλοντικοί Παράγοντες

2.1.2.1 <u>Μέσος Όρος Θερμοκρασίας</u>

Statistics		
AVG_TEMP		
N	Valid	26256
	Missing	15
Mean		13,859
Median		14,600
Std. Deviation		9,5203
Variance		90,637
Skewness		-,194
Std. Error of Skewness		,015
Kurtosis		-,942
Std. Error of Kurtosis		,030
Minimum		-14,8
Maximum		33,9

Πίνακας 2.8.Περιγραφικά μέτρα της μέσης Θερμοκρασίας

Από το ιστόγραμμα ο συντελεστής ασυμμετρίας (skewness=-0,194) έχει αρνητικό πρόσημο δηλαδή έχουμε ελαφρώς αρνητική ασυμμετρία που σημαίνει ότι η εξάπλωση της ασθένειας κυμαίνεται γύρω από τις μέσες θερμοκρασίες μεταξύ 20 και 30 βαθμών κελσίου όπως επίσης ότι η κατανομή μας είναι πλατύκυρτη (kurtosis=-0,0942<3) που σημαίνει ότι η κυρτότητα των μέσων θερμοκρασιών είναι μικρότερη σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής και άρα συγκέντρωση τιμών γύρω από το μέσο.



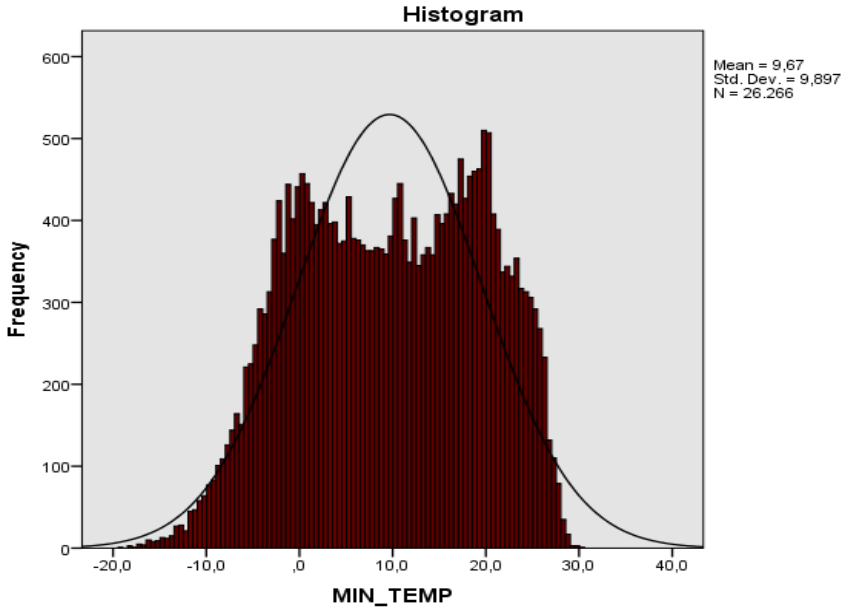
Διάγραμμα 2.8.Διακύμανση της μέσης Θερμοκρασίας

2.1.2.2 Ελάχιστη Θερμοκρασία

Statistics		
MIN_TEMP		
N	Valid	26266
	Missing	5
Mean		9,665
Median		9,900
Std. Deviation		9,8968
Variance		97,946
Skewness		-,112
Std. Error of Skewness		,015
Kurtosis		-,988
Std. Error of Kurtosis		,030
Minimum		-19,2
Maximum		30,3

Πίνακας 2.9.Περιγραφικά μέτρα ελάχιστης Θερμοκρασίας

Παρατηρούμε ότι στο ιστόγραμμα ο συντελεστής ασυμμετρίας (skewness=-0,112) έχει αρνητικό πρόσημο δηλαδή έχουμε ελαφρώς αρνητική ασυμμετρία που σημαίνει ότι η εξάπλωση της ασθένειας κυμαίνεται γύρω από τις ελάχιστες θερμοκρασίες, δηλαδή το μεγαλύτερο ποσοστό κυμαίνεται από -10 έως 20 βαθμούς κελσίου όπως επίσης ότι η κατανομή μας είναι πλατύκυρτη (kurtosis =-0,988<3) που σημαίνει ότι η κυρτότητα των ελάχιστων θερμοκρασιών είναι μικρότερη σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής και άρα συγκέντρωση τιμών γύρω από το μέσο.



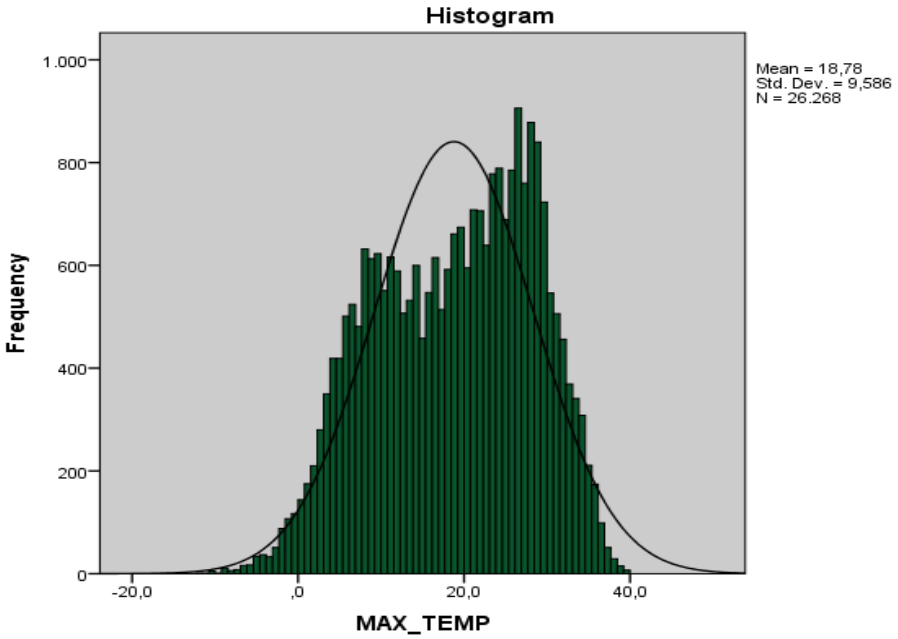
Διάγραμμα 2.9.Διακύμανση ελάχιστης Θερμοκρασίας

2.1.2.3 Μέγιστη Θερμοκρασία

Statistics		
MAX_TEMP		
N	Valid	26268
	Missing	3
Mean		18,779
Median		19,800
Std. Deviation		9,5863
Variance		91,897
Skewness		-,244
Std. Error of Skewness		,015
Kurtosis		-,859
Std. Error of Kurtosis		,030
Minimum		-11,9
Maximum		40,0

Πίνακας 2.10.Περιγραφικά μέτρα μέγιστης Θερμοκρασίας

Όσον αφορά τη μέγιστη θερμοκρασία διαφέρει από την ελάχιστη θερμοκρασία καθώς στο διάγραμμα ο συντελεστής ασυμμετρίας (skewness=-0,244) έχει αρνητικό πρόσημο δηλαδή έχουμε ελαφρώς αρνητική ασυμμετρία που σημαίνει ότι η εξάπλωση της ασθένειας κυμαίνεται γύρω από τις μέγιστες θερμοκρασίες και συγκεκριμένα στους 20 έως 40 βαθμούς κελσίου όπως επίσης ότι η κατανομή μας είναι πλατύκυρτη (kurtosis =-0,859<3) που σημαίνει ότι η κυρτότητα των μέγιστων θερμοκρασιών είναι μικρότερη σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής και άρα συγκέντρωση τιμών γύρω από το μέσο.



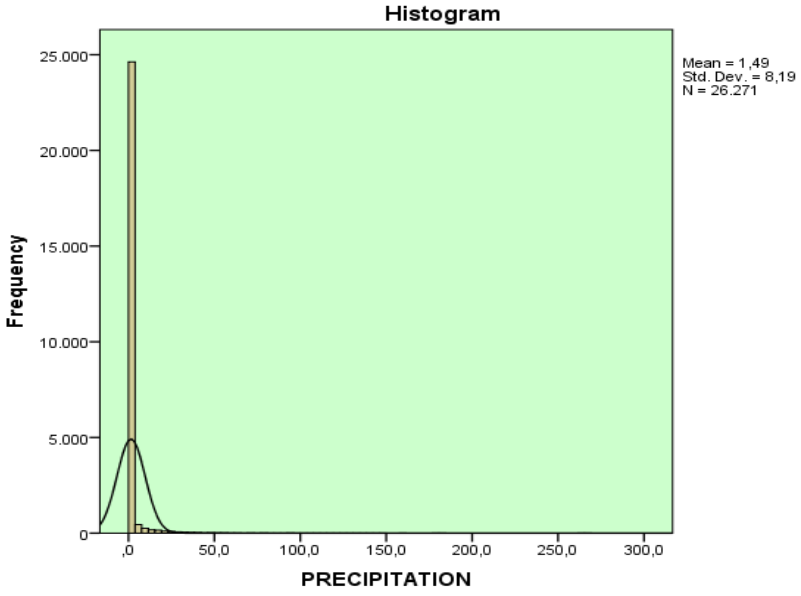
Διάγραμμα 2.10.Διακύμανση μέγιστης Θερμοκρασίας

2.1.2.4 <u>Βροχόπτωση</u>

Statistics		
PRECIPITATION		
N	Valid	26271
	Missing	0
Mean		1,487
Median		,000
Std. Deviation		8,1896
Variance		67,070
Skewness		10,917
Std. Error of Skewness		,015
Kurtosis		183,612
Std. Error of Kurtosis		,030
Minimum		,0
Maximum		266,0

Πίνακας 2.11. Περιγραφικά μέτρα Βροχόπτωσης

Ο παράγοντας βροχόπτωση δεν λαμβάνεται υπόψιν το οποίο παρατηρείται έντονα και στο παρακάτω ιστόγραμμα. Συγκεκριμένα, ο συντελεστής ασυμμετρίας (skewness=10,917) έχει θετικό πρόσημο δηλαδή έχουμε ελαφρώς θετική ασυμμετρία που σημαίνει ότι η εξάπλωση της ασθένειας επηρεάζεται ελάχιστα από τον παράγοντα βροχόπτωση όπως επίσης ότι η κατανομή μας είναι λεπτόκυρτη (kurtosis =183,612>3) που σημαίνει ότι η κυρτότητα των τιμών της βροχόπτωσης είναι μεγαλύτερες σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής χωρίς να μπορούμε να εξάγουμε κάποια ουσιαστικά συμπεράσματα για τον συγκεκριμένο περιβαλλοντικό παράγοντα.



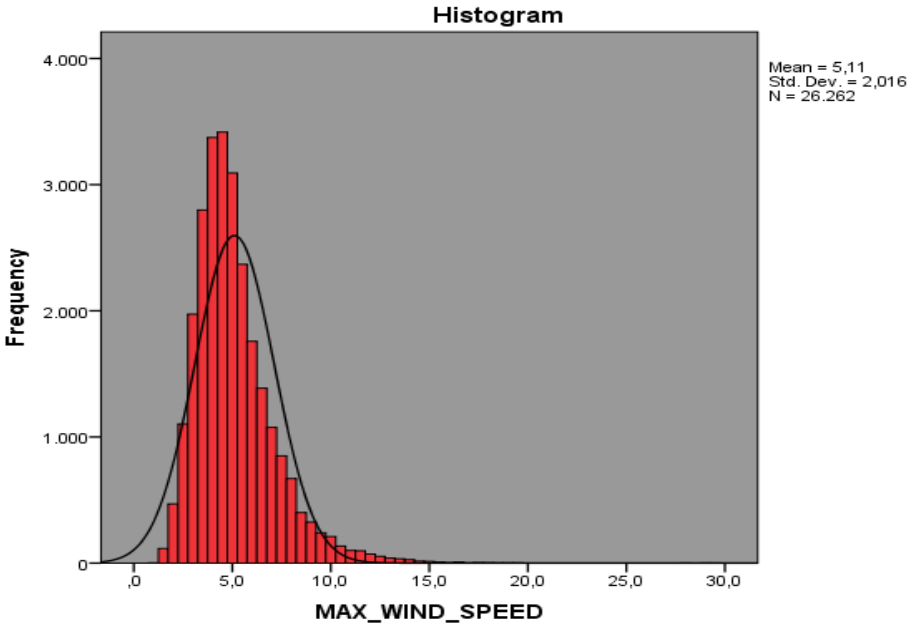
Διάγραμμα 2.11.Διακύμανση Βροχόπτωσης

2.1.2.5 Μέγιστη Ταχύτητα Ανέμου

Statistics		
MAX_WIND_SPEED		
N	Valid	26262
	Missing	9
Mean		5,110
Median		4,700
Std. Deviation		2,0163
Variance		4,065
Skewness		1,598
Std. Error of Skewness		,015
Kurtosis		5,647
Std. Error of Kurtosis		,030
Minimum		1,0
Maximum		29,4

Πίνακας 2.12. Περιγραφικά μέτρα μέγιστης ταχύτητας Ανέμου

Το ιστόγραμμα σύμφωνα με τον συντελεστή ασυμμετρίας (skewness=1,598) έχει θετικό πρόσημο δηλαδή έχουμε ελαφρώς θετική ασυμμετρία που σημαίνει ότι η εξάπλωση της ασθένειας κυμαίνεται γύρω από μέγιστες ταχύτητες ανέμου και συγκεκριμένα από τις 4.0 μέχρι 6.0 όπως επίσης ότι η κατανομή μας είναι λεπτόκυρτη (kurtosis =5,647>3) που σημαίνει ότι η κυρτότητα των μέγιστων ταχυτήτων ανέμου είναι μεγαλύτερη σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής και άρα συγκέντρωση τιμών γύρω από το μέσο.



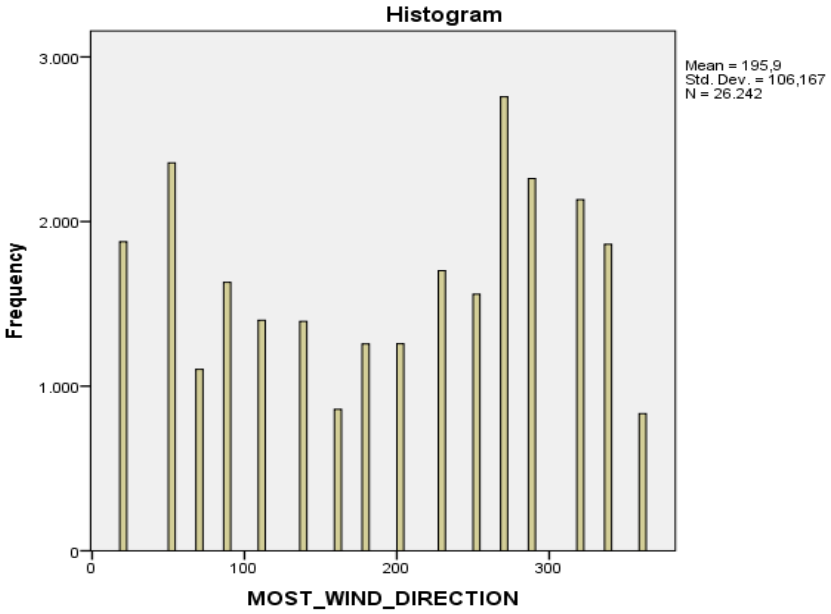
Διάγραμμα 2.12. Διακύμανση μέγιστης ταχύτητας Ανέμου

2.1.2.6 Κατεύθυνση Ανέμου

Statistics		
MOST_WIND_DIRECTION		
N	Valid	26242
	Missing	29
Mean		195,90
Median		200,00
Std. Deviation		106,167
Variance		11271,331
Skewness		-,197
Std. Error of Skewness		,015
Kurtosis		-1,312
Std. Error of Kurtosis		,030
Minimum		20
Maximum		360

Πίνακας 2.13. Περιγραφικά μέτρα κατεύθυνσης Ανέμου

Όσον αφορά την διεύθυνση του ανέμου στο ιστόγραμμα έχουμε ελαφρώς αρνητική ασυμμετρία λόγω του ότι ο συντελεστής ασυμμετρίας (skewness=-0,197) είναι αρνητικός που σημαίνει ότι η συχνότητα του ανέμου δεν είναι προς κάποια κατεύθυνση χωρίς έχουμε συνοχή ή σύνδεση. Επίσης η κατανομή μας είναι πλατύκυρτη (kurtosis =-1,312<3) που σημαίνει ότι η κυρτότητα της κατεύθυνσης του ανέμου είναι μικρότερη σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής.



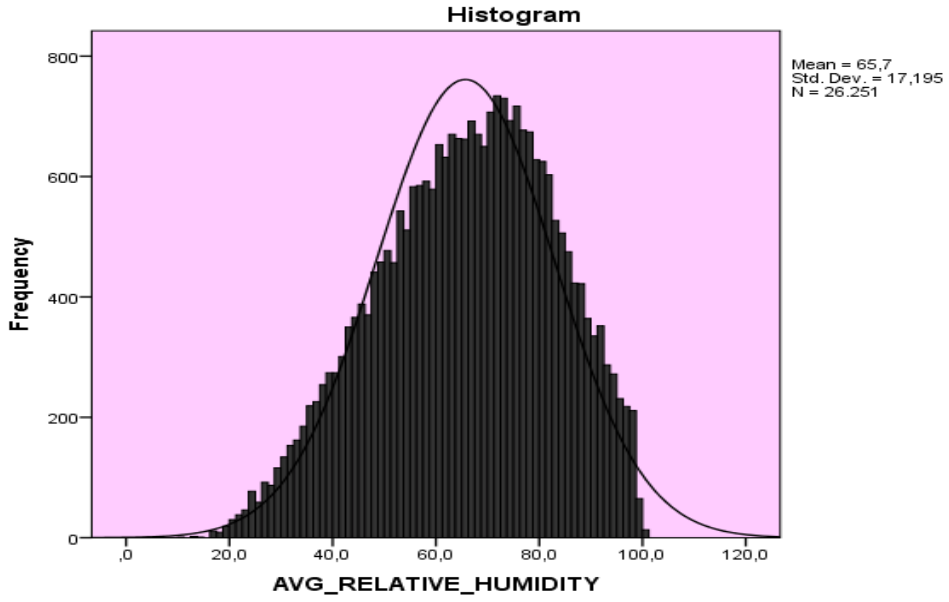
Διάγραμμα 2.13. Διακύμανση κατεύθυνσης Ανέμου

2.1.2.7 Μέση Σχετική Υγρασία

Statistics		
AVG_RELATIVE_HUMIDITY		
N	Valid	26251
	Missing	20
Mean		65,696
Median		66,900
Std. Deviation		17,1953
Variance		295,680
Skewness		-,274
Std. Error of Skewness		,015
Kurtosis		-,558
Std. Error of Kurtosis		,030
Minimum		10,4
Maximum		100,0

Πίνακας 2.14. Περιγραφικά μέτρα μέσης Υγρασίας

Για τη μέση σχετική υγρασία στο ιστόγραμμα παρατηρείται ότι ο συντελεστής ασυμμετρίας (skewness=-0,274) έχει αρνητικό πρόσημο δηλαδή έχουμε ελαφρώς αρνητική ασυμμετρία που σημαίνει ότι η εξάπλωση της ασθένειας κυμαίνεται γύρω από υψηλές τιμές όπως επίσης και ότι η κατανομή μας είναι πλατύκυρτη (kurtosis =-0,558<3) που σημαίνει ότι η κυρτότητα των τιμών της μέσης σχετικής υγρασίας είναι μικρότερη σε σχέση με την συνήθη κυρτότητα της κανονικής κατανομής και άρα συγκέντρωση τιμών γύρω από το μέσο.



Διάγραμμα 2.14.Διακύμανση μέσης Υγρασίας

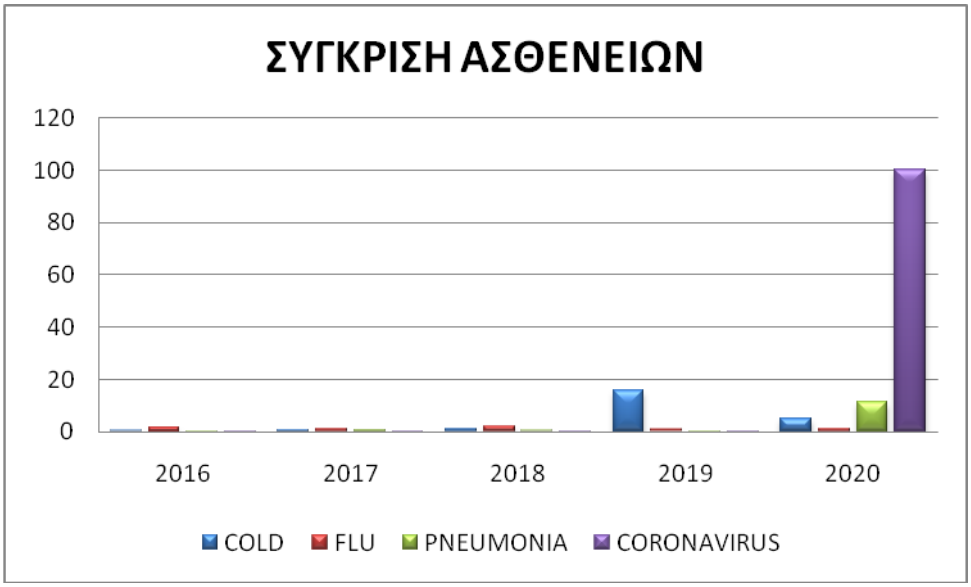
2.2 Κορωνοϊός σε σχέση με Άλλες Ασθένειες

Στην ενότητα αυτή θα ασχοληθούμε με την σύγκριση του κορωνοϊού σε σχέση με το κρυολόγημα, την πνευμονία και την γρίπη για τις χρονιές 2016-2020 ώστε να εξάγουμε κάποια αρχικά συμπεράσματα για την κατάσταση των ασθενειών στην Νότια Κορέα τα τελευταία τέσσερα χρόνια.

	COLD			FLU		
	MIN	MAX	MEAN	MIN	MAX	MEAN
2016	0,0817	0,5158	0,15184	0,0195	1,7550	0,29718
2017	0,0731	0,7808	0,13913	0,0177	1,0621	0,12303
2018	0,0880	0,9349	0,15724	0,0228	2,0714	0,23724
2019	0,0556	15,7207	0,22073	0,0283	0,8209	0,22952
2020	0,0853	5,0273	0,41562	0,0213	1,1816	0,18483
	PNEUMONIA			CORONAVIRUS		
	MIN	MAX	MEAN	MIN	MAX	MEAN
2016	0,1057	0,2904	0,19181	0,0000	0,0000	0,0000
2017	0,0913	0,5353	0,17047	0,0000	0,0000	0,0000
2018	0,0788	0,3947	0,15724	0,0000	0,0000	0,0000
2019	0,0989	0,2475	0,17007	0,0000	0,0000	0,0000
2020	0,0713	11,3932	0,62218	0,0169	100	16,80247

ΠΙΝΑΚΑΣ 2.15. Μέσο Ημερήσιο Ποσοστό των ασθενειών, 2016-2020

Από τον πίνακα των μέσων ημερήσιων ποσοστών παρατηρούμε ότι μέσα σε αυτά τα τέσσερα χρόνια η γρίπη, η πνευμονία και το κρυολόγημα είναι σε υψηλά επίπεδα, όμως παρατηρείται ότι το 2020 υπήρχε έντονη αύξηση της πνευμονίας ταυτόχρονα με την εμφάνιση του κορωνοϊού που σημαίνει ότι σχετίζονται οι δύο αυτές ασθένειες.



Διάγραμμα 2.15.Σύγκριση των τεσσάρων ασθενειών

Σύμφωνα με το παραπάνω διάγραμμα για την διάρκεια 2016-2020 παρατηρούμε ότι κατά τη διάρκεια 2016-2019 και οι τέσσερις ασθένειες δεν παρουσιάζουν ιδιαίτερη ανοδική τάση, μόνο το κρυολόγημα το 2019 φαίνεται να επηρεάζει την υγεία των πολιτών. Αντιθέτως, το 2020 ο κορωνοϊός βρίσκεται σε πολύ υψηλά επίπεδα που σημαίνει ότι η εμφάνιση της νέας ασθένειας ελαχιστοποίησε αρκετά την εξάπλωση των άλλων ασθενειών.

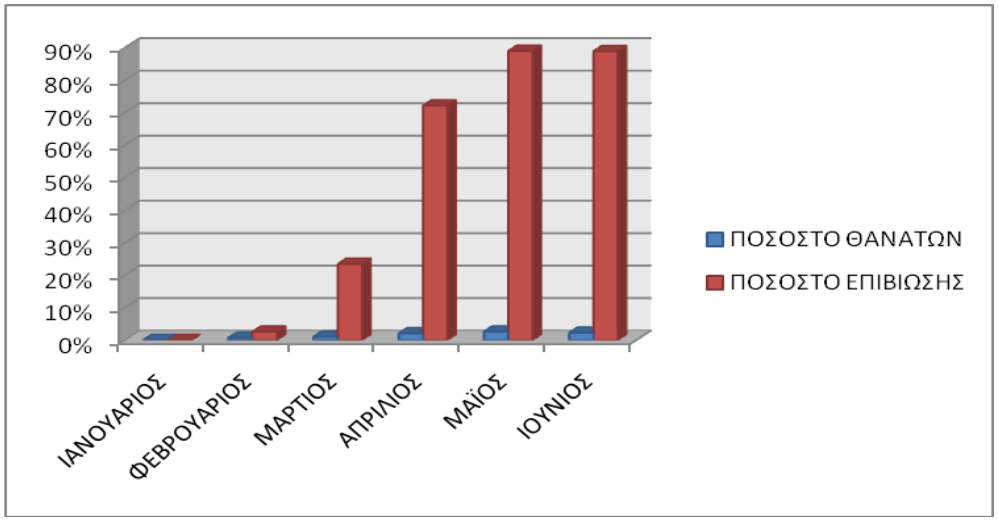
2.3 Χρονολογική Σύγκριση του Κορωνοϊού

Η ενότητα αυτή αναφέρεται στην σύγκριση της ασθένειας με βάση τον χρόνο αλλά εξειδικεύεται και στην σύγκριση με βάση την ηλικία, το φύλο και την επαρχία επηρεαζόμενη και πάλι από το χρόνο.

ΜΗΝΑΣ	ΠΟΣΟΣΤΟ ΑΡΝΗΤΙΚΩΝ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΘΕΤΙΚΩΝ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΕΠΙΒΙΩΣΗΣ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ
ΙΑΝΟΥΑΡΙΟΣ	82,74%	3,79%	0%	0%
ΦΕΒΡΟΥΑΡΙΟΣ	66,22%	2,31%	2,61%	0,71%
ΜΑΡΤΙΟΣ	89,98%	2,92%	23,47%	1,07%
ΑΠΡΙΛΙΟΣ	95,41%	1,97%	72,17%	2,07%
ΜΑΪΟΣ	96,34%	1,49%	88,98%	2,65%
ΙΟΥΝΙΟΣ	96,83%	1,09%	88,91%	2,28%

ΠΙΝΑΚΑΣ 2.16.Ποσοστά χρονολογικής σύγκρισης του Κορωνοϊού

Από τον παραπάνω πίνακα παρατηρείται ότι με την πάροδο του χρόνου το ποσοστό των αρνητικών τεστ αυξάνεται ενώ το ποσοστό των θετικών μειώνεται. Το ίδιο παρατηρείται και στα ποσοστά επιβίωσης και θανάτων αντίστοιχα που σημαίνει ότι αρκετοί πολίτες της Νοτίου Κορέας κατάφεραν να ξεπεράσουν τον κορωνοϊό, το οποίο φαίνεται και στο παρακάτω διάγραμμα. Ωστόσο τα ποσοστά θανάτων δεν τείνουν στην μονάδα καθώς η εμφάνιση της ασθένειας δεν εξαλείφθηκε τον Ιούνιο αλλά συνέχισε να εξελίσσεται.



Διάγραμμα 2.16. Ποσοστά χρονολογικής σύγκρισης του Κορωνοϊού

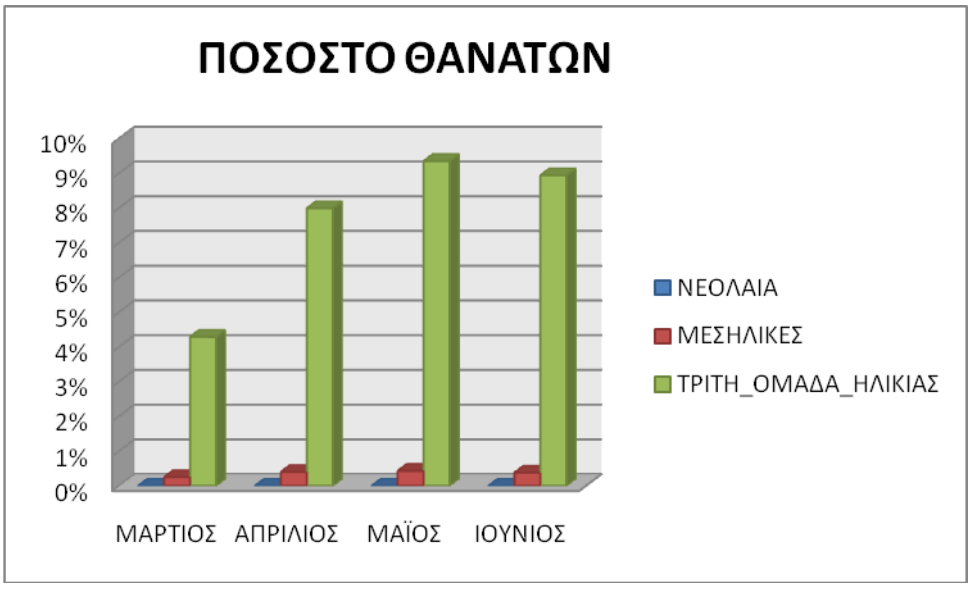
2.3.1 Χρονολογική Σύγκριση με βάση την Ηλικία

Για να συγκρίνουμε τον κορωνοϊό με βάση την ηλικία ταξινομήσαμε τις ηλικίες σε τρείς κατηγορίες:

- Από 0 έως 29 ετών ως νεολαία
- Από 30 έως 59 ετών ως μεσήλικες
- Από 60 έως 89 ετών ως τρίτη_ομάδα_ηλικίας

	ΝΕΟΛΑΙΑ		ΜΕΣΗΛΙΚΕΣ		ΤΡΙΤΗ_ΟΜΑΔΑ_ΗΛΙΚΙΑΣ	
ΜΗΝΑΣ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ
ΜΑΡΤΙΟΣ	81756	0%	103724	0,25%	53677	4,29%
ΑΠΡΙΛΙΟΣ	107080	0%	133395	0,40%	74883	7,99%
ΜΑΪΟΣ	118420	0%	144067	0,43%	79743	9,35%
ΙΟΥΝΙΟΣ	123430	0%	154633	0,38%	86395	8,94%

ΠΙΝΑΚΑΣ 2.17. Ποσοστά χρονολογικής σύγκρισης ανά Ηλικία



Διάγραμμα 2.17. Ποσοστά χρονολογικής σύγκρισης ανά Ηλικία

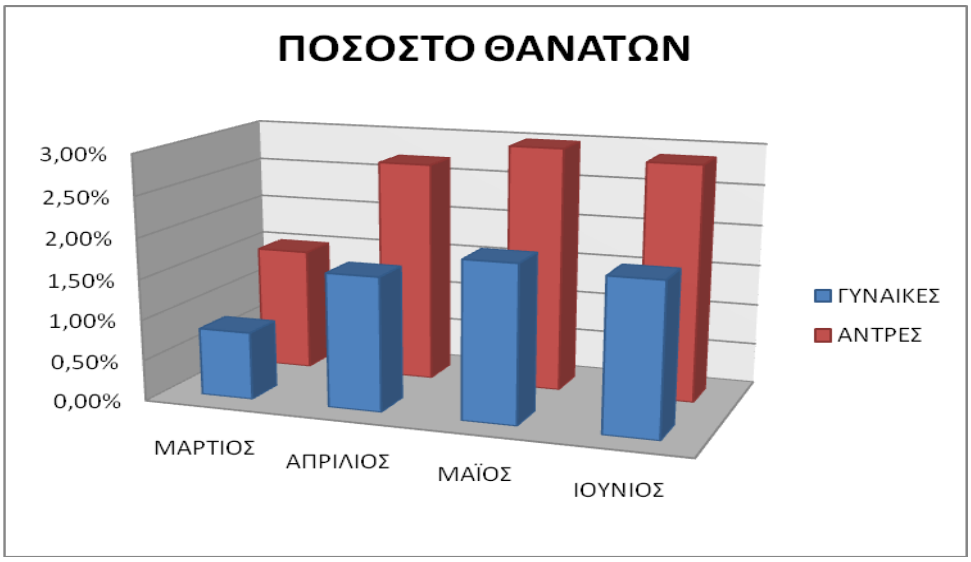
Από το ιστόγραμμα συμπεραίνουμε ότι η συγκεκριμένη ασθένεια επηρέασε περισσότερο τα άτομα της τρίτης ηλικιακής ομάδας καθώς είχαμε αυξημένους θανάτους με το πέρασμα του χρόνου ενώ στις νεότερες ηλικίες είχαμε μηδαμινούς θανάτους από τον ιό. Με άλλα λόγια τους συγκεκριμένους μήνες ο Covid-19 είχε μεγαλύτερη επίπτωση στους ηλικιωμένους παρά στους νεότερους ανθρώπους όσον αφορά τους ασθενείς της Νότιας Κορέας.

2.3.2 Χρονολογική Σύγκριση με βάση το Φύλο

	ΓΥΝΑΙΚΕΣ		ΑΝΤΡΕΣ	
ΜΗΝΑΣ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ
ΜΑΡΤΙΟΣ	147103	0,83%	92044	1,51%
ΑΠΡΙΛΙΟΣ	188504	1,64%	126855	2,71%
ΜΑΪΟΣ	201558	1,92%	140672	2,99%
ΙΟΥΝΙΟΣ	210302	1,85%	154156	2,88%

ΠΙΝΑΚΑΣ 2.18. Ποσοστά χρονολογικής σύγκρισης ανά Φύλο

Όσον αφορά τους άντρες και τις γυναίκες αυτό που παρατηρείται είναι ότι παρόλο που στις γυναίκες έχουμε περισσότερα θετικά τεστ σε σχέση με τους άνδρες, το ποσοστό θανάτων είναι μεγαλύτερο στους άντρες παρά στις γυναίκες για τους μήνες που πραγματοποιείται η ανάλυσή μας, κάτι το οποίο μπορούμε να παρατηρήσουμε και στο παρακάτω ιστόγραμμα.



Διάγραμμα 2.18. Ποσοστά χρονολογικής σύγκρισης ανά Φύλο

2.3.3 Χρονολογική Σύγκριση με βάση την Επαρχία

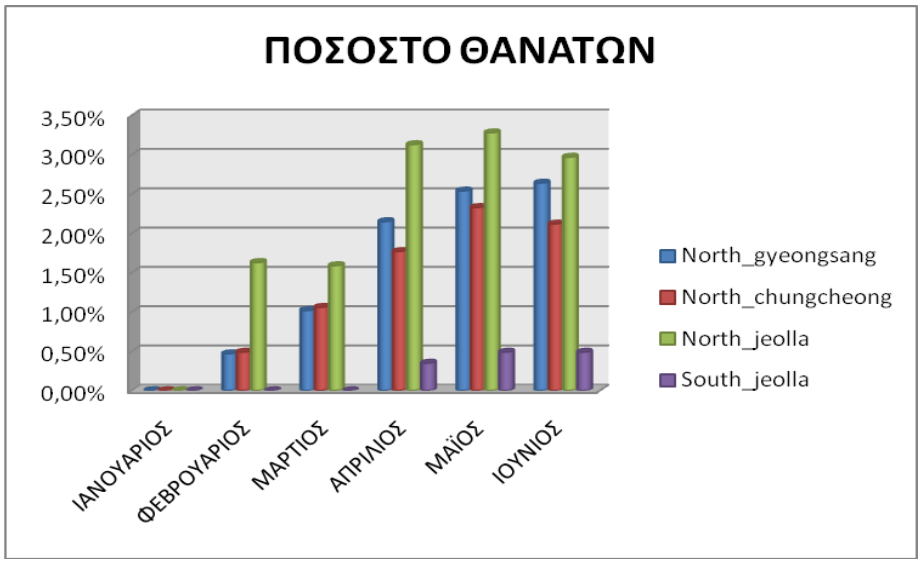
Για να συγκρίνουμε τον κορωνοϊό με βάση την επαρχία ταξινομήσαμε τις επαρχίες σε τέσσερις κατηγορίες:

- North_gyeongsang
- North_chungcheong
- North_jeolla
- South_jeolla

	North_gyeongsang			North_chungcheong		
ΜΗΝΑΣ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΕΠΙΒΙΩΣΗΣ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΕΠΙΒΙΩΣΗΣ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ
ΙΑΝΟΥΑΡΙΟΣ	0	0%	0%	18	0%	0%
ΦΕΒΡΟΥΑΡΙΟΣ	7883	0,12%	0,47%	801	17,48%	0,49%
ΜΑΡΤΙΟΣ	183590	23,23%	1,02%	9974	25,27%	1,06%
ΑΠΡΙΛΙΟΣ	213756	79,49%	2,15%	22264	54,55%	1,77%
ΜΑΪΟΣ	223426	92,85%	2,54%	26528	81,57%	2,33%
ΙΟΥΝΙΟΣ	217773	96,47%	2,64%	35958	71,69%	2,12%
	North_jeolla			South_jeolla		
ΜΗΝΑΣ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΕΠΙΒΙΩΣΗΣ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ	ΘΕΤΙΚΑ ΤΕΣΤ	ΠΟΣΟΣΤΟ ΕΠΙΒΙΩΣΗΣ	ΠΟΣΟΣΤΟ ΘΑΝΑΤΩΝ
ΙΑΝΟΥΑΡΙΟΣ	12	0%	0%	11	0%	0%
ΦΕΒΡΟΥΑΡΙΟΣ	2700	1,63%	1,63%	769	14,95%	0%
ΜΑΡΤΙΟΣ	38119	1,59%	1,59%	9822	21,62%	0%
ΑΠΡΙΛΙΟΣ	46627	3,13%	3,13%	21462	46,17%	0,35%
ΜΑΪΟΣ	50569	3,28%	3,28%	26444	79,37%	0,49%
ΙΟΥΝΙΟΣ	55006	2,97%	2,97%	37666	68,37%	0,49%

ΠΙΝΑΚΑΣ 2.19. Ποσοστά χρονολογικής σύγκρισης ανά Επαρχία

Από τον παραπάνω πίνακα παρατηρούμε ότι και στις τέσσερις κατηγορίες είναι αυξημένο το ποσοστό θνησιμότητας ενώ έχουμε αρκετά μικρά ποσοστά θανάτων. Αξιοσημείωτο είναι , ωστόσο ότι στην κατηγορία North_Jeolla τα ποσοστά επιβίωσης και θανάτων αντίστοιχα είναι αρκετά μικρά όπως επίσης στην κατηγορία South_Jeolla το ποσοστό αποθανόντων είναι σχεδόν μηδενικό.



Διάγραμμα 2.19. Ποσοστά χρονολογικής σύγκρισης ανά Επαρχία

ΚΕΦΑΛΑΙΟ 3^0

3.1 Εισαγωγή

Στο κεφάλαιο αυτό θα χρησιμοποιήσουμε τις μεταβλητές από τους δημογραφικούς παράγοντες που θεωρούμε ότι είναι σημαντικές για να εξάγουμε συμπεράσματα σχετικά με την εξάπλωση του κορωνοϊού στην Νότια Κορέα. Θα συγκρίνουμε τους παράγοντες αυτούς με την κατάσταση υγείας των πολιτών της χώρας ώστε να έχουμε μια πλήρη εικόνα με το πόσο έχει επηρεάσει ο καινούργιος ιός τη Νότια Κορέα.

3.1.1 Συσχέτιση με τη μεταβλητή "Φύλο"

Σε αυτό το σημείο θα ασχοληθούμε με την ύπαρξη συσχέτισης ανάμεσα στο φύλο των ασθενών και την κατάσταση υγείας τους, δηλαδή αν απεβίωσε, παραμένει ακόμα στο νοσοκομείο ή επέστρεψε στην κατοικία του αφού ξεπέρασε τον κίνδυνο. Παρατηρούμε ότι κανένα κελί του πίνακα συνάφειας δεν έχει αναμενόμενη συχνότητα μικρότερη του 5 και έτσι ισχύουν οι προϋποθέσεις του ελέγχου χ^2 του Pearson.

ΦΥΛΟ * ΚΑΤΑΣΤΑΣΗ Crosstabulation					
Count					
		ΚΑΤΑΣΤΑΣΗ			Total
		deceased	released	isolated	
ΦΥΛΟ	male	47	1112	666	1825
	female	28	1402	787	2217
Total		75	2514	1453	4042

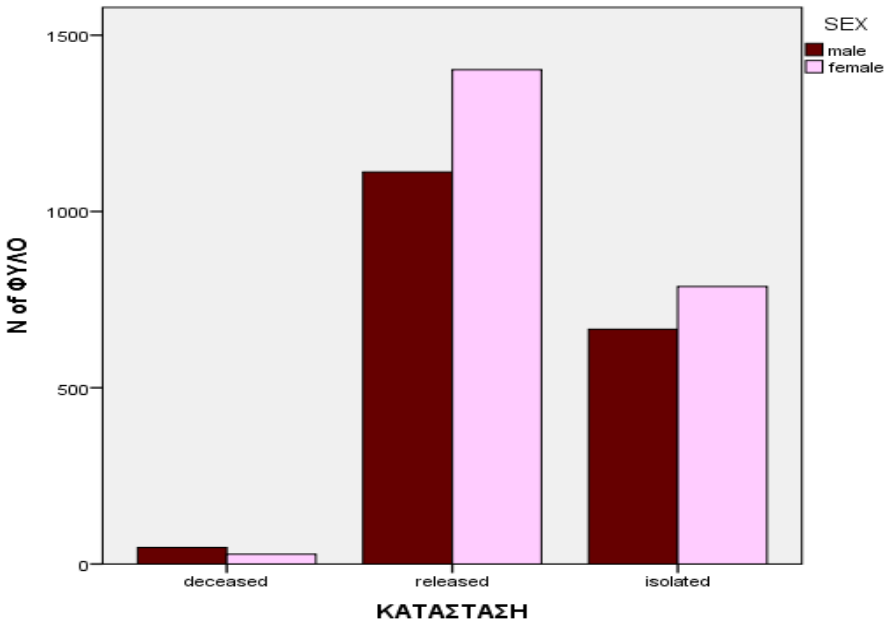
Πίνακας 3.1.Πίνακας συνάφειας ανά Φύλο

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,424[a]	2	,005
Likelihood Ratio	10,405	2	,006
Linear-by-Linear Association	,039	1	,844
N of Valid Cases	4042		
a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 33,86.			

Πίνακας 3.2.Έλεγχος χ^2 του Pearson ανά Φύλο

Παρατηρούμε ότι η στατιστική συνάρτηση του ελέγχου χ^2 του Pearson ισούται με 10,424 και για επίπεδο σημαντικότητας 5% το p-value=0,005 είναι μικρότερο που

σημαίνει ότι υπάρχει εξάρτηση μεταξύ του φύλου και της κατάστασης του ασθενούς, δηλαδή είναι στατιστικά σημαντική η εξάρτηση των δύο μεταβλητών. Σύμφωνα με το παρακάτω διάγραμμα διακρίνουμε ότι οι άνδρες ασθενείς είναι αυτοί που αποβιώνουν περισσότερο σε σχέση με τις γυναίκες ωστόσο η διαφορά είναι μικρή ενώ στις περιπτώσεις όπου ο ασθενής παίρνει εισιτήριο από το νοσοκομείο ή παραμένει σε αυτό φαίνεται πιο ξεκάθαρα ότι μεγαλύτερο ποσοστό γυναικών ανήκουν σε αυτές τις δύο τις κατηγορίες.



Διάγραμμα 3.1.Κατάταξη της συσχέτισης ανά Φύλο

3.1.2 Συσχέτιση με τη μεταβλητή "Ηλικία"

Σε αυτό το σημείο θα ασχοληθούμε με την ύπαρξη συσχέτισης ανάμεσα στην ηλικία των ασθενών και την κατάσταση υγείας τους, δηλαδή αν απεβίωσε, παραμένει ακόμα στο νοσοκομείο ή επέστρεψε στην κατοικία του αφού ξεπέρασε τον κίνδυνο. Θα χρησιμοποιήσουμε τον μη παραμετρικό έλεγχο Kruskal-Wallis καθώς οι υποθέσεις περί κανονικότητας των πληθυσμών της ηλικίας των ασθενών δεν μπορούν να επαληθευτούν.

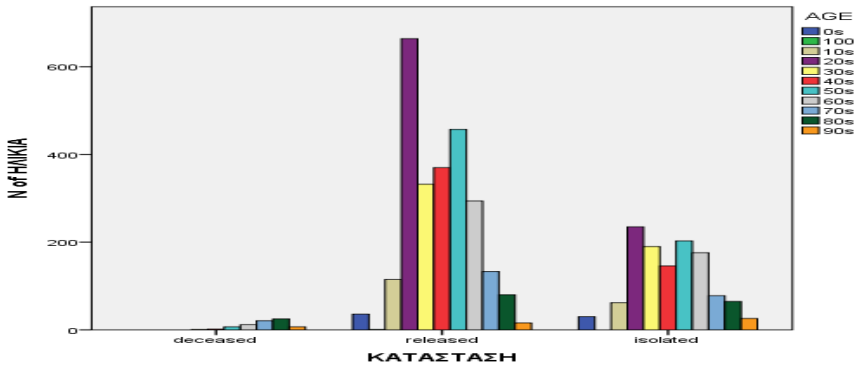
Ranks			
	ΗΛΙΚΙΑ	N	Mean Rank
ΚΑΤΑΣΤΑΣΗ	0s	66	2167,45
	10s	177	1974,10
	20s	899	1809,27
	30s	523	1995,76
	40s	518	1842,23
	50s	667	1875,41
	60s	482	1969,63
	70s	232	1831,55
	80s	170	1844,38
	90s	49	2124,73
	100s	1	1324,50
	Total	3784	

Πίνακας 3.3.Τα σκορ ανά Ηλικία

Test Statistics[a,b]	
	ΚΑΤΑΣΤΑΣΗ
Chi-Square	32,754
df	10
Asymp. Sig.	,000
a. Kruskal Wallis Test	
b. Grouping Variable: ΗΛΙΚΙΑ	

Πίνακας 3.4.Έλεγχος Kruskal-Wallis ανά Ηλικία

Η τιμή της συνάρτησης Η των Kruskal-Wallis ισούται με 32,754 και το p-value είναι 0,0001, δηλαδή για επίπεδο σημαντικότητας α=5% παρατηρούμε ότι διαφέρει η κατάσταση του κάθε ασθενή ανάλογα με την ηλικία του που σημαίνει ότι υπάρχει εξάρτηση μεταξύ των δύο μεταβλητών. Σύμφωνα με το παρακάτω διάγραμμα μπορούμε να συμπεράνουμε ότι τα άτομα ηλικίας 50 ετών και άνω είναι αυτά που αποβιώνουν λόγω του κορωνοϊού, άτομα ηλικίας 20 ετών μέχρι 50 ετών καταφέρνουν να αναρρώσουν από την συγκεκριμένη ασθένεια ενώ φαίνεται να είναι λίγα τα άτομα ηλικίας 20 ετών μέχρι 60 ετών που παραμένουν ακόμα στο νοσοκομείο.



Διάγραμμα 3.2. Κατάταξη της συσχέτισης ανά Ηλικία

3.1.3 Συσχέτιση με τη μεταβλητή "Επαρχία"

Σε αυτό το σημείο θα ασχοληθούμε με την ύπαρξη συσχέτισης ανάμεσα στην επαρχία στην οποία κατοικούν οι ασθενείς και την κατάσταση υγείας τους, δηλαδή αν απεβίωσε, παραμένει ακόμα στο νοσοκομείο ή επέστρεψε στην κατοικία του αφού ξεπέρασε τον κίνδυνο. Παρατηρούμε ότι έντεκα κελιά του πίνακα συνάφειας που αντιστοιχεί 24,4% του συνόλου των κελιών, έχουν αναμενόμενη συχνότητα μικρότερη του 5 και έτσι παραβιάζεται μία από τις προϋποθέσεις του ελέγχου χ^2 του Pearson και για αυτό λόγο θα χρησιμοποιήσουμε την προσομοίωση Monte-Carlo.

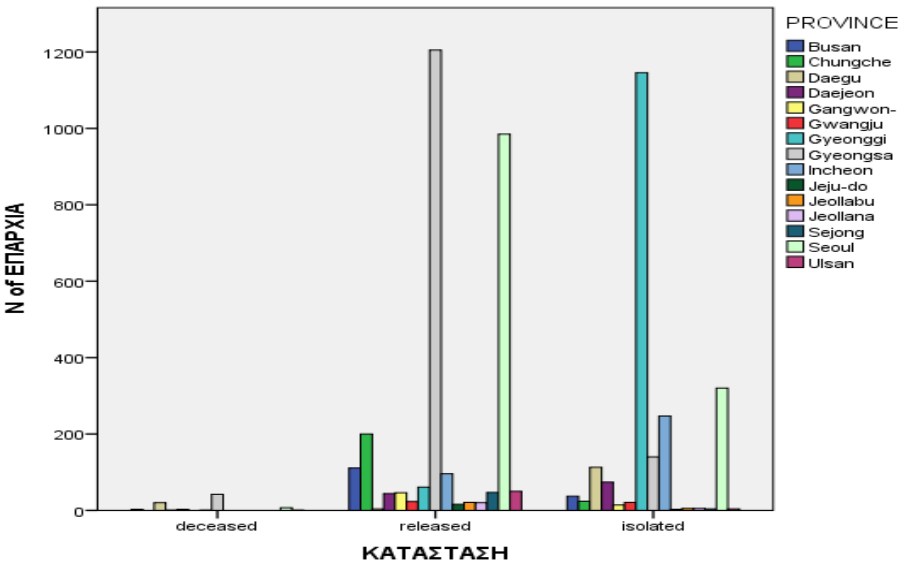
ΕΠΑΡΧΙΑ * ΚΑΤΑΣΤΑΣΗ Crosstabulation					
Count					
		ΚΑΤΑΣΤΑΣΗ			Total
		deceased	released	isolated	
ΕΠΑΡΧΙΑ	Busan	3	111	37	151
	Chungche	0	200	24	224
	Daegu	20	4	113	137
	Daejeon	1	44	74	119
	Gangwon-	3	46	14	63
	Gwangju	0	23	21	44
	Gyeonggi	1	61	1146	1208
	Gyeongsa	42	1205	140	1387
	Incheon	0	96	247	343
	Jeju-do	0	16	3	19
	Jeollabu	0	21	5	26
	Jeollana	0	20	5	25
	Sejong	0	47	4	51
	Seoul	7	985	320	1312
	Ulsan	1	50	4	55
Total		78	2929	2157	5164

Πίνακας 3.5.Πίνακας συνάφειας ανά Επαρχία

Chi-Square Tests									
	Value	df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)			Monte Carlo Sig. (1-sided)		
				Sig.	99% Confidence Interval		Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound		Lower Bound	Upper Bound
Pearson Chi-Square	2789,043[a]	28	,000	,000[b]	,000	,000			
Likelihood Ratio	3071,653	28	,000	,000[b]	,000	,000			
Fisher's Exact Test	3050,030			,000[b]	,000	,000			
Linear-by-Linear Association	161,376[c]	1	,000	,000[b]	,000	,000	,000[b]	,000	,000
N of Valid Cases	5164								
a. 11 cells (24,4%) have expected count less than 5. The minimum expected count is ,29.									
b. Based on 10000 sampled tables with starting seed 624387341.									
c. The standardized statistic is -12,703.									

Πίνακας 3.6.Έλεγχος Fisher's ανά Επαρχία

Αφού δεν ισχύει ο έλεγχος χ^2 του Pearson χρησιμοποιούμε τον έλεγχο Fisher's Exact Test όπου η στατιστική συνάρτηση ισούται με 3050,030 και για επίπεδο σημαντικότητας 5% το p-value=0,000 είναι μικρότερο που σημαίνει ότι υπάρχει εξάρτηση μεταξύ της επαρχίας στην οποία κατοικεί ο κάθε ασθενείς σε σχέση με την κατάστασή του, δηλαδή είναι στατιστικά σημαντική η εξάρτηση των δύο μεταβλητών. Σύμφωνα με το παρακάτω διάγραμμα παρατηρούμε ότι δεν υπάρχει κάποια συγκεκριμένη επαρχία από την οποία να απεβίωσαν οι περισσότεροι ασθενείς, όσον αφορά την ανάρρωση των ασθενών και την αποχώρησή τους από το νοσοκομείο φαίνεται ότι οι περισσότεροι προέρχονται από τις επαρχίες Jeollana και Seoul ενώ μεγάλο ποσοστό ασθενών που παραμένουν στο νοσοκομείο είναι από την επαρχία Gyeonggi.



Διάγραμμα 3.3. Κατάταξη της συσχέτισης ανά Επαρχία

3.2 Λογιστική Παλινδρόμηση με τις μεταβλητές "Φύλο", "Ηλικία" και " Επαρχία"

Σύμφωνα με τους πίνακες συνάφειας παρατηρήσαμε ότι υπάρχει συσχέτιση μεταξύ φύλου, ηλικίας και επαρχίας με την κατάσταση του κάθε ασθενούς, ωστόσο θέλουμε να διευρύνουμε την ανάλυσή μας και να ελέγξουμε κατά πόσο οι τρεις αυτές μεταβλητές επηρεάζουν την κατάσταση υγείας του κάθε ατόμου. Θα χρησιμοποιήσουμε το μοντέλο λογιστικής παλινδρόμησης καθώς η κατάσταση του κάθε ασθενή είναι δίτιμη μεταβλητή όπου παίρνει την τιμή μηδέν όταν ο ασθενής αποβιώνει και την τιμή ένα όταν ο ασθενής αναρρώνει και αποχωρεί από το νοσοκομείο.

Στο αρχείο αποτελεσμάτων ο πίνακας Dependent Variable Encoding μας δίνει πληροφορίες σχετικά με την κωδικοποίηση της μεταβλητής απόκρισης του μοντέλου μας ενώ ο πίνακας Categorical Variables Encoding μας πληροφορεί για τις συχνότητες και την κωδικοποίηση των ανεξάρτητων μεταβλητών που θα χρησιμοποιήσουμε.

Dependent Variable Encoding	
Original Value	Internal Value
deceased	0
released	1

Πίνακας 3.7. Κωδικοποίηση μεταβλητής απόκρισης

Categorical Variables Codings				
		Frequency	Parameter coding	
			(1)	(2)
CODE_PROVINCE	Gyeongsang	432	1,000	,000
	Jeolla	1440	,000	1,000
	Chungcheong	717	,000	,000
NEW_AGE	Νεολαία	1832	1,000	,000
	Μεσήλικες	565	,000	1,000
	Τρίτη_Ομάδα_Ηλικίας	192	,000	,000
NEW_SEX	male	1159	1,000	
	female	1430	,000	

Πίνακας 3.8. Κωδικοποίηση ανεξάρτητων μεταβλητών

Στον πίνακα Model Summary πληροφορούμαστε σχετικά με το ποσοστό ερμηνείας της μεταβλητότητας του μοντέλου μας. Συγκεκριμένα το Cox & Snell R Square ισούται με 0,059 και το Nagelkerke R Square με 0,255 που σημαίνει ότι το μοντέλο πολλαπλής παλινδρόμησης που προσαρμόσαμε ερμηνεύει το 6% με 30% περίπου της μεταβλητότητας.

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	521,882[a]	,059	,255
a. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.			

Πίνακας 3.9. Ποσοστό μεταβλητότητας του μοντέλου

Από τον πίνακα Hosmer and Lemeshow Test περιλαμβάνονται τα αποτελέσματα για την καλή προσαρμογή του μοντέλου μας, δηλαδή κατά πόσο κοντά βρίσκονται οι παρατηρούμενες και οι προβλεπόμενες πιθανότητες. Από το αποτέλεσμα του ελέγχου προκύπτει ότι το μοντέλο μας προσαρμόζεται αρκετά ικανοποιητικά στα δεδομένα, καθώς το p-value=0,445 που είναι μεγαλύτερο από το επίπεδο σημαντικότητας α=5%.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	6,843	7	,445

Πίνακας 3.10. Έλεγχος καλής προσαρμογής

Σύμφωνα με τον πίνακα ταξινόμησης το ποσοστό ορθής ταξινόμησης ισούται με 97,1% που σημαίνει ότι η προσθήκη των συγκεκριμένων ερμηνευτικών μεταβλητών οδήγησε στην αύξηση της προβλεπτικής αξίας του μοντέλου.

Classification Table[a]					
	Observed		Predicted		
			ΚΑΤΑΣΤΑΣΗ_ΥΓΕΙΑΣ		Percentage Correct
			deceased	released	
Step 1	ΚΑΤΑΣΤΑΣΗ_ΥΓΕΙΑΣ	deceased	0	75	,0
		released	0	2514	100,0
	Overall Percentage				97,1
a. The cut value is ,500					

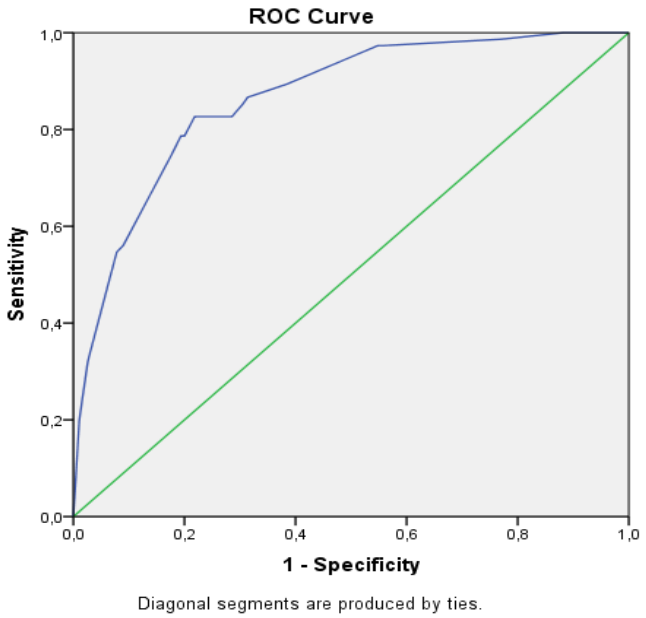
Πίνακας 3.11.Πίνακας Ταξινόμησης

Το μοντέλο αν και έχει ποσοστό ορθής πρόβλεψης 97% εν τούτοις δίνει την τιμή 1 σε όλες τις περιπτώσεις προβλέποντας ότι όλοι θα ξεπεράσουν το πρόβλημα. Αυτό συμβαίνει γιατί οι θάνατοι είναι πολύ λίγοι σε σχέση με αυτούς που έγιναν καλά. Συγκεκριμένα, θα χρησιμοποιήσουμε την ανάλυση ROC στην οποία θα εντοπίσουμε το βέλτιστο σημείο αποκοπής ώστε να το εφαρμόσουμε στην λογιστική παλινδρόμηση και να εξάγουμε πιο ουσιώδη αποτελέσματα.

Case Processing Summary	
ΚΑΤΑΣΤΑΣΗ_ΥΓΕΙΑΣ	Valid N (listwise)
Positive[a]	75
Negative	2514
Missing	2576
Smaller values of the test result variable(s) indicate stronger evidence for a positive actual state.	
a. The positive actual state is deceased.	

Πίνακας 3.12.Διαχωρισμός θετικών και αρνητικών περιπτώσεων

Από τον πίνακα Case Processing Summary πληροφορούμαστε ότι έχουμε 75 θετικές καταστάσεις οι οποίες αναφέρονται στους ασθενείς της Νότιας Κορέας ενώ 2514 αρνητικές καταστάσεις αναφέρονται στους υγιείς.



Διάγραμμα 3.4. Καμπύλη ROC Curve

Σύμφωνα με την καμπύλη ROC παρατηρούμε ότι η μπλε «καμπύλη» βρίσκεται αρκετά κοντά στον κάθετο άξονα που σημαίνει ότι οι εκτιμήσεις του συγκεκριμένου μοντέλου λογιστικής παλινδρόμησης έχουν υψηλή διαγνωστική αξία.

Area Under the Curve				
Test Result Variable(s): Predicted probability				
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,865	,020	,000	,826	,904
The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.				
a. Under the nonparametric assumption				
b. Null hypothesis: true area = 0.5				

Πίνακας 3.13.Πίνακας αξιολόγησης της περιοχής κάτω από την καμπύλη

Για να αξιολογήσουμε τη διαγνωστική αξία των εκτιμήσεων κοιτάμε την περιοχή κάτω από την καμπύλη. Σύμφωνα με τον πίνακα η περιοχή κάτω από την καμπύλη ισούται με 0,865 η οποία είναι αρκετά υψηλή τιμή. Η μηδενική υπόθεση ότι η περιοχή κάτω από την καμπύλη ισούται με 0,5 απορρίπτεται καθώς το p-value του ελέγχου είναι αρκετά μικρότερο από το 0,05 που σημαίνει ότι οι εκτιμητές του συγκεκριμένου μοντέλου αποτελεί έναν αρκετά αξιόπιστο διαγνωστικό έλεγχο.

Coordinates of the Curve		
Test Result Variable(s): Predicted probability		
Positive if Less Than or Equal To[a]	Sensitivity	1 - Specificity
0E-7	,000	,000
,7777138	,200	,011
,8562306	,320	,026
,8811992	,547	,078
,9190004	,560	,089
,9429014	,747	,176
,9523950	,787	,194
,9578200	,787	,200
,9687361	,827	,218
,9789312	,827	,238
,9813222	,827	,286
,9843617	,853	,305
,9887724	,867	,314
,9918711	,893	,383
,9931656	,973	,548
,9954253	,973	,557
,9970519	,987	,771
,9981339	1,000	,883
1,0000000	1,000	1,000
The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group.		
a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.		

Πίνακας 3.14.Πίνακας επιλογής σημείου αποκοπής

Από το διάγραμμα για την καμπύλη η διαγώνιος θεωρείται το τυχαίο μοντέλο με πιθανότητα επιτυχίας 0,05 και παίρνουμε το σημείο που είναι πιο μακριά από το τυχαίο μοντέλο το οποίο ισούται με 0,96. Αυτό σημαίνει ότι με βάση τους εκτιμητές αν ένα άτομο έχει τιμή μικρότερη ή ίση από 0,96 τότε θεωρείται ασθενής, διαφορετικά θεωρείται υγιής.

Σύμφωνα με την ανάλυση ROC για να εφαρμόσουμε την λογιστική παλινδρόμηση στα δεδομένα μας θα χρησιμοποιήσουμε το σημείο αποκοπής που ισούται με 0,96 ώστε να εξάγουμε πιο ουσιώδη συμπεράσματα.

Στο αρχείο αποτελεσμάτων ο πίνακας Dependent Variable Encoding μας δίνει πληροφορίες σχετικά με την κωδικοποίηση της μεταβλητής απόκρισης του μοντέλου μας ενώ ο πίνακας Categorical Variables Encoding μας πληροφορεί για τις συχνότητες και την κωδικοποίηση των ανεξάρτητων μεταβλητών που θα χρησιμοποιήσουμε.

Dependent Variable Encoding	
Original Value	Internal Value
deceased	0
released	1

Πίνακας 3.15. Κωδικοποίηση μεταβλητής απόκρισης

Categorical Variables Codings				
		Frequency	Parameter coding	
			(1)	(2)
CODE_PROVINCE	Gyeongsang	432	1,000	,000
	Jeolla	1440	,000	1,000
	Chungcheong	717	,000	,000
NEW_AGE	Νεολαία	1832	1,000	,000
	Μεσήλικες	565	,000	1,000
	Τρίτη_Ομάδα_Ηλικίας	192	,000	,000
NEW_SEX	male	1159	1,000	
	female	1430	,000	

Πίνακας 3.16. Κωδικοποίηση ανεξάρτητων μεταβλητών

Στον πίνακα Model Summary πληροφορούμαστε σχετικά με το ποσοστό ερμηνείας της μεταβλητότητας του μοντέλου μας. Συγκεκριμένα το Cox & Snell R Square ισούται με 0,059 και το Nagelkerke R Square με 0,255 που σημαίνει ότι το μοντέλο πολλαπλής παλινδρόμησης που προσαρμόσαμε ερμηνεύει το 6% με 30% περίπου της μεταβλητότητας.

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	521,882[a]	,059	,255
a. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.			

Πίνακας 3.17. Ποσοστό μεταβλητότητας του μοντέλου

Από τον πίνακα Hosmer and Lemeshow Test περιέχει τα αποτελέσματα για την καλή προσαρμογή του μοντέλου μας, δηλαδή κατά πόσο κοντά βρίσκονται οι παρατηρούμενες και οι προβλεπόμενες πιθανότητες. Από το αποτέλεσμα του ελέγχου προκύπτει ότι το μοντέλο μας προσαρμόζεται αρκετά ικανοποιητικά στα δεδομένα, καθώς το p-value=0,445 που είναι μεγαλύτερο από το επίπεδο σημαντικότητας α=5%.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	6,843	7	,445

Πίνακας 3.18. Έλεγχος καλής προσαρμογής

Σύμφωνα με τον πίνακα ταξινόμησης το ποσοστό ορθής ταξινόμησης ισούται με 78,3% που σημαίνει ότι η προσθήκη των συγκεκριμένων ερμηνευτικών μεταβλητών οδήγησε στην αύξηση της προβλεπτικής αξίας του μοντέλου.

Classification Table[a]					
	Observed		Predicted		
			ΚΑΤΑΣΤΑΣΗ_ΥΓΕΙΑΣ		Percentage Correct
			deceased	released	
Step 1	ΚΑΤΑΣΤΑΣΗ_ΥΓΕΙΑΣ	deceased	62	13	82,7
		released	549	1965	78,2
	Overall Percentage				78,3
a. The cut value is ,960					

Πίνακας 3.19.Πίνακας Ταξινόμησης

Στον πίνακα Variables in the Equation έχουμε το προσαρμοσμένο μοντέλο το οποίο μας δίνει πληροφορίες για την σχέση της μεταβλητής απόκρισης με τις ανεξάρτητες μεταβλητές. Από τις τιμές των p-value κρίνουμε ποιες μεταβλητές είναι στατιστικά σημαντικές και διαπιστώνουμε ότι και οι τρεις μεταβλητές φύλο, ηλικία και επαρχία είναι στατιστικά σημαντικές στο μοντέλο μας για επίπεδο σημαντικότητας α=5%.

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	NEW_SEX(1)	-,873	,254	11,837	1	,001	,417	,254	,687
	NEW_AGE			78,015	2	,000			
	NEW_AGE(1)	1,682	,505	11,090	1	,001	5,374	1,997	14,457
	NEW_AGE(2)	-1,339	,421	10,112	1	,001	,262	,115	,598
	CODE_PROVINCE			21,898	2	,000			
	CODE_PROVINCE(1)	-2,049	,503	16,581	1	,000	,129	,048	,346
	CODE_PROVINCE(2)	-1,090	,483	5,079	1	,024	,336	,130	,868
	Constant	5,129	,626	67,139	1	,000	168,924		
a. Variable(s) entered on step 1: NEW_SEX, NEW_AGE, CODE_PROVINCE.									

Πίνακας 3.20.Εκτίμηση των παραμέτρων του μοντέλου

Το μοντέλο πολλαπλής λογιστικής παλινδρόμησης που θα χρησιμοποιήσουμε είναι το εξής:

$$\text{Log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{NEW\_SEX}(1) + \beta_2 \text{NEW\_AGE}(1) + \beta_3 \text{NEW\_AGE}(2) + \beta_4 \text{CODE\_PROVINCE}(1) + \beta_5 \text{CODE\_PROVINCE}(2)$$

$$\text{Log}\left(\frac{p}{1-p}\right) = 5.129 - 0.873\text{NEW\_SEX}(1) + 1.682\text{NEW\_AGE}(1) - 1.339\text{NEW\_AGE}(2) - 2.049\text{CODE\_PROVINCE}(1) - 1.090\text{CODE\_PROVINCE}(2)$$

$$P = \frac{exp(5.129 - 0.873NEW\_SEX(1) + 1.682NEW\_AGE(1) - 1.339NEW\_AGE(2) - 2.049CODE\_PROVINCE(1) - 1.090CODE\_PROVINCE(2))}{1 + exp(5.129 - 0.873NEW\_SEX(1) + 1.682NEW\_AGE(1) - 1.339NEW\_AGE(2) - 2.049CODE\_PROVINCE(1) - 1.090CODE\_PROVINCE(2))}$$

Ο σχετικός λόγος πιθανοτήτων για τις γυναίκες ισούται με 0,417 που σημαίνει ότι η πιθανότητα μια γυναίκα ασθενής να επιβιώσει είναι μειωμένη κατά 1-41,7%=58,3% σε σχέση με έναν άνδρα ασθενή, όταν οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Ο σχετικός λόγος πιθανοτήτων για τους μεσήλικες ισούται με 5,374 που σημαίνει ότι ένα άτομο που κατατάσσεται στην κατηγορία μεσήλικες είναι 5,374 φορές πιο πιθανόν να επιβιώση σε σχέση με ένα άτομο που κατατάσσεται στην κατηγορία νεολαία, όταν οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Ο σχετικός λόγος πιθανοτήτων για τα άτομα που ανήκουν στην τρίτη_ομάδα_ηλικίας ισούται με 0,262 που σημαίνει ότι η πιθανότητα επιβίωσης ενός ασθενή που ανήκει στη τρίτη_ομάδα_ηλικίας είναι μειωμένη κατά 1-26,2%=73,8% σε σχέση με έναν ασθενή που ανήκει στη νεολαία, όταν οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

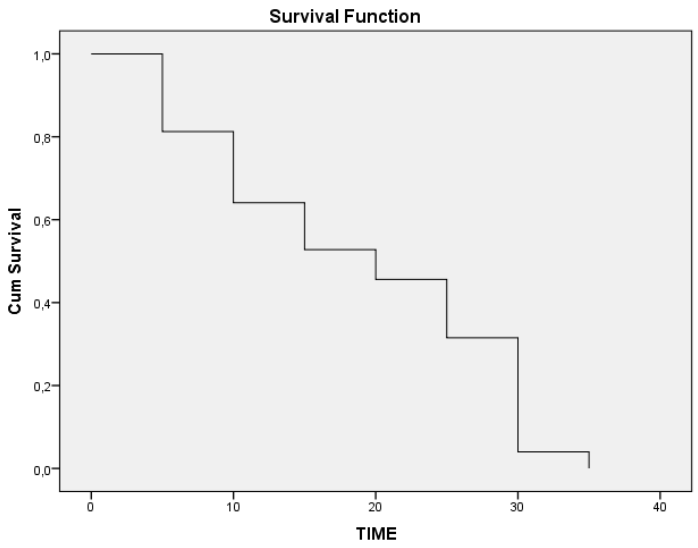
Ο σχετικός λόγος πιθανοτήτων για την ευρύτερη περιοχή Jeolla ισούται με 0,129 που σημαίνει ότι η πιθανότητα επιβίωσης ενός ατόμου που ανήκει στην ευρύτερη περιοχή Jeolla είναι μειωμένη κατά 1-12,9%=87,1% σε σχέση με ένα άτομο που ανήκει στην ευρύτερη περιοχή Gyeongsang, όταν οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Ο σχετικός λόγος πιθανοτήτων για την ευρύτερη περιοχή Chungcheong ισούται με 0,336 που σημαίνει ότι η πιθανότητα επιβίωσης ενός ασθενή που βρίσκεται στην ευρύτερη περιοχή Chungcheong είναι μειωμένη κατά 1-33,6%=66,4% σε σχέση με έναν ασθενή που ανήκει στην ευρύτερη περιοχή Gyeongsang, όταν οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

3.3 Πίνακες Επιβίωσης

Στην ενότητα αυτή θα ασχοληθούμε με τους πίνακες επιβίωσης που είναι από τους πρώτους και πιο διαδεδομένους μεθόδους στη δημογραφία και στον αναλογισμό για την περιγραφή δεδομένων σχετικά με χρόνους επιβίωσης. Χρησιμοποιούνται για την αναπαράσταση του ποσοστού επιβίωσης μιας μεγάλης ομάδας ατόμων που παρακολουθείται στο χρόνο. Στη συγκεκριμένη μελέτη θα ασχοληθούμε με διαγράμματα επιβίωσης για το σύνολο των ασθενών αλλά και με βάση το φύλο, την ηλικία και την επαρχία από την οποία προέρχονται τα άτομα αλλά θα χρησιμοποιήσουμε και την μέθοδο Kaplan-Meier.

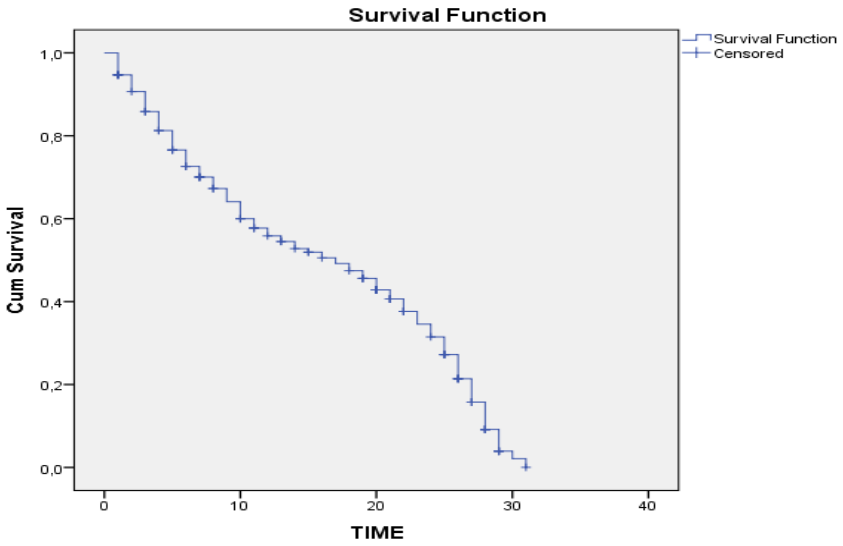
3.3.1 Διάγραμμα Επιβίωσης για το σύνολο των ατόμων



Διάγραμμα 3.5.Συνάρτηση επιβίωσης των ασθενών

Σύμφωνα με το διάγραμμα επιβίωσης παρατηρούμε ότι στο διάστημα [30,39) το τελευταίο άτομο της ανάλυσης μας επιβιώνει από την ασθένεια και αποχωρεί από το νοσοκομείο.

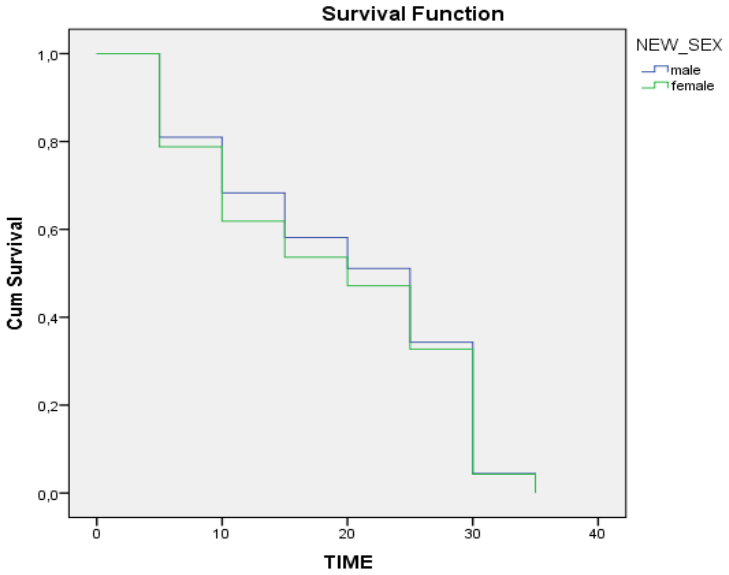
3.3.1.1 Διάγραμμα Επιβίωσης για το σύνολο των ατόμων με την μέθοδο Kaplan-Meier



Διάγραμμα 3.6.Συνάρτηση επιβίωσης των ασθενών με την μέθοδο Kaplan-Meier

Από την γραφική παράσταση προκύπτει ότι στο σύνολο των ασθενών η ασθένεια έχει σημαντική επίδραση με αποτέλεσμα να υπάρχουν αρκετοί θάνατοι.

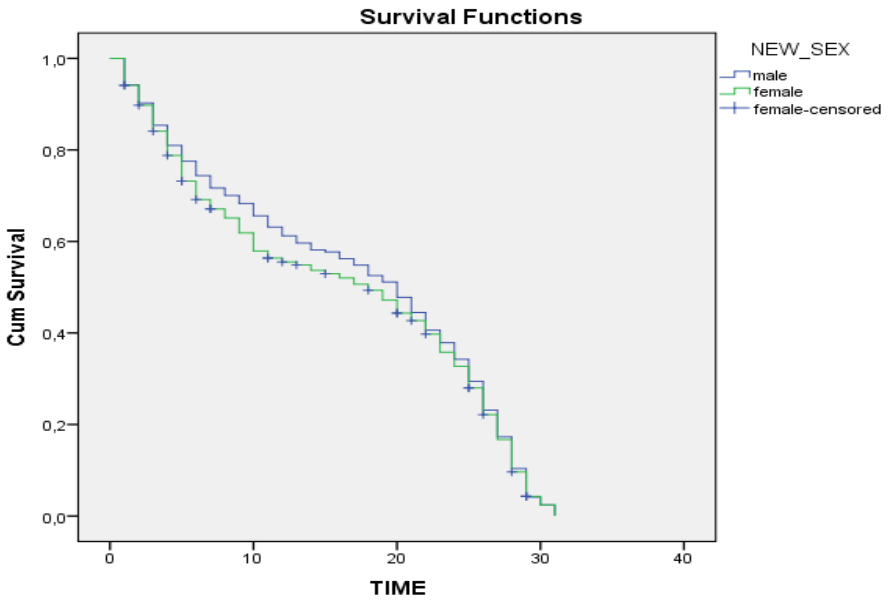
3.3.2 Διάγραμμα Επιβίωσης για τη μεταβλητή "Φύλο"



Διάγραμμα 3.7. Συνάρτηση επιβίωσης των ασθενών ανά Φύλο

Από το διάγραμμα επιβίωσης διαπιστώνουμε ότι κανένας ασθενής είτε άντρας είτε γυναίκα αποβιώνουν από τον κορωνοϊό στο διάστημα [30,39), αντιθέτως αναρρώνουν και αποχωρούν από το νοσοκομείο.

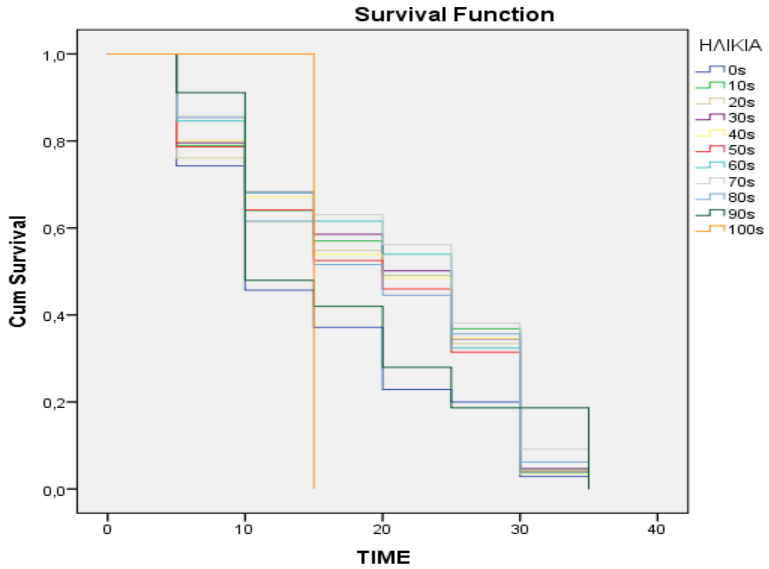
3.3.2.1 Διάγραμμα Επιβίωσης για τη μεταβλητή ‘‘Φύλο’’ με την μέθοδο Kaplan-Meier



Διάγραμμα 3.8. Συνάρτηση επιβίωσης των ασθενών ανά Φύλο με την μέθοδο Kaplan-Meier

Από την γραφική παράσταση προκύπτει ότι στις γυναίκες η επίδραση του κορωνοϊού είναι μεγαλύτερη σε σχέση με τους άνδρες καθώς εμφανίζεται με ταχύτατο ρυθμό στις γυναίκες.

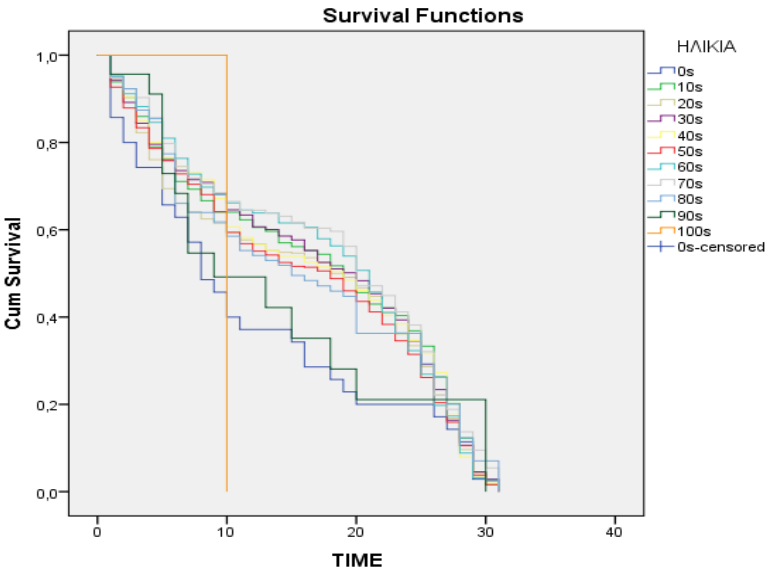
3.3.3 Διάγραμμα Επιβίωσης για τη μεταβλητή “Ηλικία”



Διάγραμμα 3.9. Συνάρτηση επιβίωσης των ασθενών ανά Ηλικία

Από το διάγραμμα επιβίωσης διαπιστώνουμε ότι στο διάστημα [30,39) όσοι ασθενείς έχουν μείνει αναρρώνουν πλήρως οποιασδήποτε ηλικίας και αποχωρούν από το νοσοκομείο.

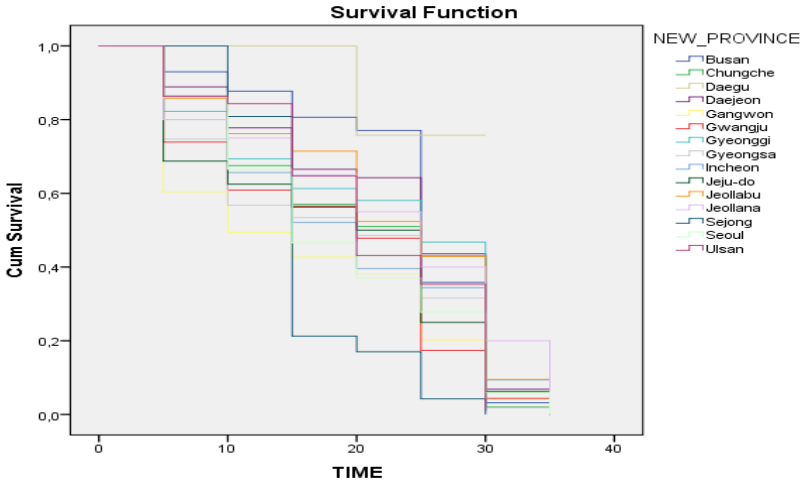
3.3.3.1 Διάγραμμα Επιβίωσης για τη μεταβλητή ‘‘Ηλικία’’ με την μέθοδο Kaplan-Meier



Διάγραμμα 3.10. Συνάρτηση επιβίωσης των ασθενών ανά Ηλικία με την μέθοδο Kaplan-Meier

Από την γραφική παράσταση παρατηρούμε ότι στα άτομα ηλικίας 100 ετών και άνω ο κορωνοϊός δεν είχε επίδραση στην κατάσταση υγείας τους ενώ στα άτομα ηλικίας 20 έως 89 ετών φαίνεται ότι η συγκεκριμένη συμβάλλει σημαντικά στην επιδείνωση της υγείας τους όπως επίσης στα άτομα 0 ετών και άνω και 90 ετών και άνω αντίστοιχα, η ασθένεια να καθυστερεί την εμφάνισή της.

3.3.4 Διάγραμμα Επιβίωσης για τη μεταβλητή “Επαρχία”

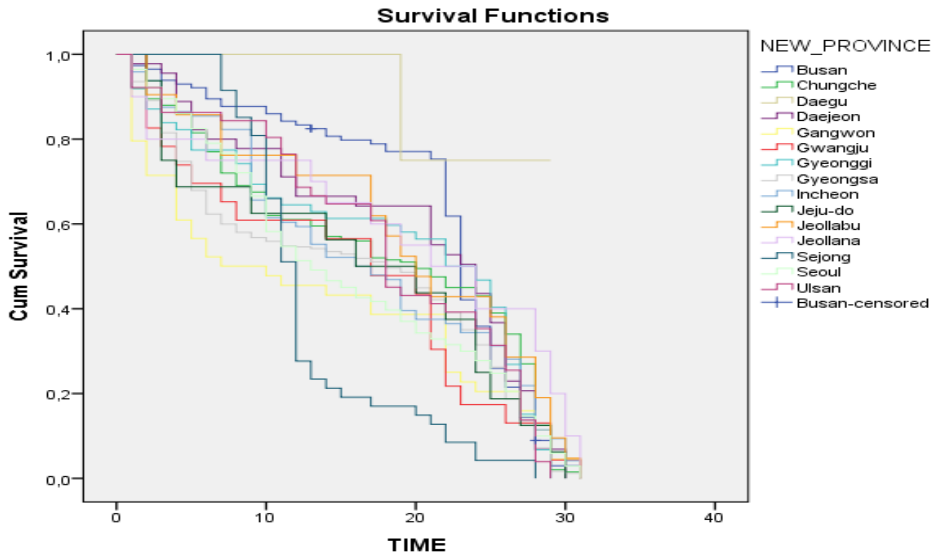


Διάγραμμα 3.11. Συνάρτηση επιβίωσης των ασθενών ανά Επαρχία

Σύμφωνα με το διάγραμμα επιβίωσης παρατηρούμε ότι οι ασθενείς όλων των επαρχιών επιβιώνουν από τον κορωνοϊό στο διάστημα [30,39) εκτός από την

επαρχία Daegu όπου στο διάστημα [26,39) τα άτομα αποβιώνουν λόγω της ασθένειας.

3.3.4.1 Διάγραμμα Επιβίωσης για τη μεταβλητή ‘‘Επαρχία’’ με την μέθοδο Kaplan-Meier



Διάγραμμα 3.12. Συνάρτηση επιβίωσης των ασθενών ανά Επαρχία με την μέθοδο Kaplan-Meier

Από την γραφική παράσταση συμπεραίνεται ότι στην επαρχία Daegu κανένα άτομο δεν απεβίωσε από το COVID-19, αντιθέτως σε αρκετές άλλες επαρχίες όπως η Jeollana και Jeollabu η εμφάνιση της ασθένειας αυτής είναι αρκετά έντονη ώστε πολλοί ασθενείς να χάνουν τη μάχη και να αποβιώνουν.

3.4 Το μοντέλο αναλογικού κινδύνου του Cox με τις μεταβλητές “Φύλο”, “Ηλικία” και ” Επαρχία”

Σε αυτό το σημείο θα ασχοληθούμε με το μοντέλο του αναλογικού κινδύνου του Cox καθώς στην έρευνά μας κρίνεται σημαντικό ο καθορισμός των συμμεταβλητών που επηρεάζουν τη συνάρτηση κινδύνου και συνεπώς την εξάπλωση του κορωνοϊού στην Νότια Κορέα την χρονική περίοδο 1/01/2020-30/06/2020.

Αρχικά παρουσιάζεται ένας πίνακας που μας δίνει γενικές πληροφορίες για τα δεδομένα μας. Συγκεκριμένα ότι 2511 από τις 2586 περιπτώσεις έχουν εμφανίσει το ενδεχόμενο ενώ οι υπόλοιπες 75 περιπτώσεις είναι λογοκριμένοι χρόνοι επιβίωσης αλλά υπάρχουν και 2579 ελλείπουσες τιμές.

Case Processing Summary			
		N	Percent
Cases available in analysis	Event[a]	2511	48,6%
	Censored	75	1,5%
	Total	2586	50,1%
Cases dropped	Cases with missing values	2579	49,9%
	Cases with negative time	0	0,0%
	Censored cases before the earliest event in a stratum	0	0,0%
	Total	2579	49,9%
Total		5165	100,0%
a. Dependent Variable: TIME			

Πίνακας 3.21.Γενικές πληροφορίες δεδομένων

Από τον πίνακα Omnibus Tests of Model Coefficients ελέγχουμε την σημαντικότητα του ολικού μοντέλου μέσω των ελέγχων score test και LR test. Σύμφωνα με τους ελέγχους αυτούς το p-value είναι αρκετά μικρό για επίπεδο σημαντικότητας α=5% και συμπεραίνουμε ότι τουλάχιστον μια από τις ερμηνευτικές μεταβλητές επηρεάζει σημαντικά τον χρόνο επιβίωσης, δηλαδή ότι το μοντέλο μας προσαρμόζεται καλά.

Omnibus Tests of Model Coefficients[a]									
-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
34862,323	74,047	26	,000	82,913	26	,000	82,913	26	,000
a. Beginning Block Number 1. Method = Enter									

Πίνακας 3.22.Έλεγχος σημαντικότητας του μοντέλου

Στον πίνακα Variables in the Equation προσαρμόζεται το μοντέλο του αναλογικού κινδύνου του Cox και σύμφωνα με τα p-value μπορούμε να κρίνουμε ποιες ερμηνευτικές μεταβλητές είναι στατιστικά σημαντικές. Παρατηρούμε ότι το φύλο και η ηλικία δεν είναι στατιστικά σημαντικές μεταβλητές αντιθέτως με την επαρχία η οποία κρίνεται σημαντική για την εξάπλωση του ιού την χρονική περίοδο 1/01/2020-1/06/2020 στην Νότια Κορέα.

Variables in the Equation								
	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
NEW_SEX	-,049	,041	1,430	1	,232	,952	,879	1,032
AGE			7,745	11	,736			
AGE(1)	-,319	,352	,817	1	,366	,727	,365	1,451
AGE(2)	-,039	,306	,016	1	,898	,962	,528	1,753
AGE(3)	,304	1,032	,087	1	,768	1,355	,179	10,240
AGE(4)	-,362	,272	1,777	1	,183	,696	,409	1,186
AGE(5)	-,294	,257	1,313	1	,252	,745	,451	1,232
AGE(6)	-,339	,260	1,698	1	,193	,712	,427	1,187
AGE(7)	-,334	,260	1,655	1	,198	,716	,430	1,191
AGE(8)	-,291	,258	1,272	1	,259	,747	,451	1,240
AGE(9)	-,361	,260	1,927	1	,165	,697	,418	1,160
AGE(10)	-,422	,268	2,478	1	,115	,656	,387	1,109
AGE(11)	-,337	,277	1,478	1	,224	,714	,415	1,229
NEW_PROVINCE			60,460	14	,000			
NEW_PROVINCE(1)	-,200	,171	1,368	1	,242	,819	,585	1,145
NEW_PROVINCE(2)	-,140	,160	,764	1	,382	,869	,635	1,190
NEW_PROVINCE(3)	-1,849	,522	12,542	1	,000	,157	,057	,438
NEW_PROVINCE(4)	-,134	,207	,420	1	,517	,874	,583	1,312
NEW_PROVINCE(5)	,198	,206	,925	1	,336	1,219	,814	1,826
NEW_PROVINCE(6)	,107	,253	,177	1	,674	1,112	,677	1,827
NEW_PROVINCE(7)	-,056	,191	,084	1	,772	,946	,650	1,376
NEW_PROVINCE(8)	,096	,145	,436	1	,509	1,100	,828	1,462
NEW_PROVINCE(9)	-,110	,176	,394	1	,530	,896	,635	1,264
NEW_PROVINCE(10)	,023	,296	,006	1	,939	1,023	,573	1,826
NEW_PROVINCE(11)	-,245	,261	,880	1	,348	,783	,470	1,306
NEW_PROVINCE(12)	-,469	,268	3,070	1	,080	,625	,370	1,057
NEW_PROVINCE(13)	,508	,206	6,109	1	,013	1,663	1,111	2,489
NEW_PROVINCE(14)	-,111	,148	,562	1	,453	,895	,669	1,196

Πίνακας 3.23.Εκτίμηση παραμέτρων του μοντέλου

Τα παραπάνω αποτελέσματα μπορούν να χρησιμοποιηθούν στην περίπτωση που θέλουμε να συμπερασματολογήσουμε για το λόγο κινδύνου δύο διαφορετικών επιπέδων, όποια και να είναι αυτά.

Ενδεικτικά,

$h\,(t|Z_2 = 0, Z_3 = 1,\ldots, Z_{15} = 0) / h\,(t|Z_2 = 1, Z_3 = 0,\ldots, Z_{15} = 0) =$
$h_0(t)\exp(\beta_3) / h_0(t)\exp(\beta_2) = \exp(\beta_3 - \beta_2)$

Ο εκτιμητής μεγίστης πιθανοφάνειας για την ποσότητα $\exp(\beta_3 - \beta_2) = 0{,}05$ που σημαίνει ότι ένας ασθενής που προέρχεται από την επαρχία Daegu έχει 0,05 φορές μεγαλύτερο κίνδυνο να αποβιώσει σε σχέση με έναν ασθενή που προέρχεται από την επαρχία Chungche.

$h\,(t|Z_2 = 0, Z_3 = 0,\ldots, Z_{15} = 1) / h\,(t|Z_2 = 1, Z_3 = 0,\ldots, Z_{14}=1, Z_{15} = 0) =$
$h_0(t)\exp(\beta_{15}) / h_0(t)\exp(\beta_{14}) = \exp(\beta_{15} - \beta_{14})$

Ο εκτιμητής μεγίστης πιθανοφάνειας για την ποσότητα $\exp(\beta_{15} - \beta_{14}) = -0{,}768$ που σημαίνει ότι ένας ασθενής που προέρχεται από την επαρχία Ulsan έχει 0,768 φορές μικρότερο κίνδυνο να αποβιώσει σε σχέση με έναν ασθενή που προέρχεται από την επαρχία Seoul.

ΚΕΦΑΛΑΙΟ 4^0

4.1 Εισαγωγή

Στο κεφάλαιο αυτό θα ασχοληθούμε για τους περιβαλλοντικούς παράγοντες που μπορούν να επηρεάσουν τη διάδοση της ασθένειας στους πολίτες της Νοτίου Κορέας. Ωστόσο, στο αρχείο μας έχουμε στοιχεία για αρκετές επαρχίες και για αυτό το λόγο θα αναφερθούμε στην Seoul ενδεικτικά. Θα αναλύσουμε χρονικά κάποιους από τους παράγοντες αυτούς καθώς και να προβλέψουμε μελλοντικές τιμές των χρονοσειρών.

4.1.1 Συσχέτιση με τις μεταβλητές "Μέσος Όρος Θερμοκρασίας", ''Μέση Σχετική Υγρασία''

Σε αυτή την ενότητα θα ασχοληθούμε με την ύπαρξη γραμμικής συσχέτισης μεταξύ των κρουσμάτων και του μέσου όρου θερμοκρασίας καθώς και της μέση σχετικής υγρασίας στην Seoul. Καθώς τα δεδομένα και των τριών μεταβλητών δεν προέρχονται από κανονική κατανομή δεν μπορούμε να χρησιμοποιήσουμε τον παραμετρικό συντελεστή του Pearson αλλά θα χρησιμοποιήσουμε τον μη παραμετρικό συντελεστή του Spearman.

	ΜΕΣΟΣ ΟΡΟΣ ΘΕΡΜΟΚΙΡΑΣΙΑΣ	ΜΕΣΗ ΣΧΕΤΙΚΗ ΥΓΡΑΣΙΑ
ΚΡΟΥΣΜΑΤΑ	CORRELATION COEFFICIENT SPEARMAN	CORRELATION COEFFICIENT SPEARMAN
10 ΗΜΕΡΕΣ	0,367	0,314
11 ΗΜΕΡΕΣ	0,347	0,226
12 ΗΜΕΡΕΣ	0,323	0,206
13 ΗΜΕΡΕΣ	0,268	0,231
14 ΗΜΕΡΕΣ	0,215	0,229
15 ΗΜΕΡΕΣ	0,162	0,205

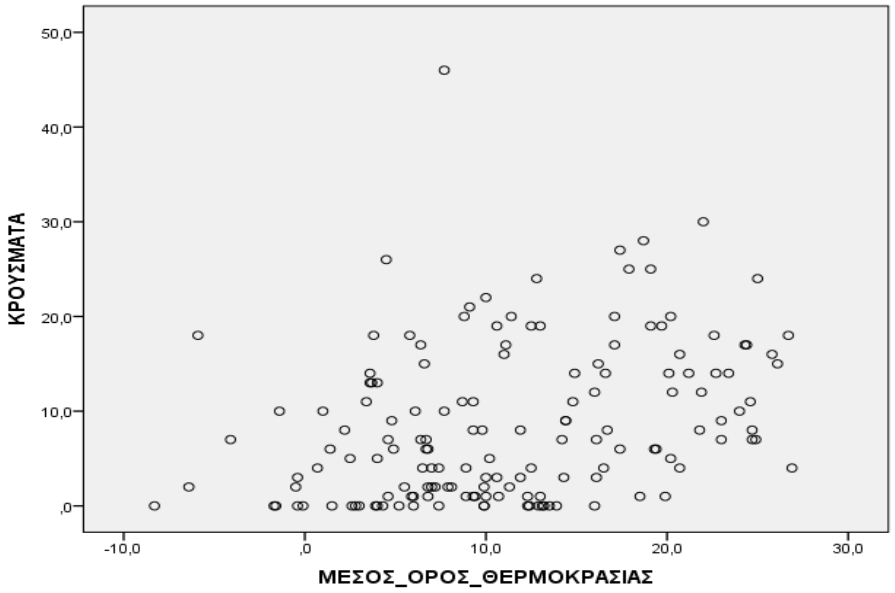
Πίνακας 4.1.Συσχέτιση κρουσμάτων με μέση Θερμοκρασία και μέση Υγρασία

Σύμφωνα με τον παραπάνω πίνακα καλύτερα αποτελέσματα όσον αφορά τα κρούσματα σε σχέση με τη μέση θερμοκρασία και τη μέση υγρασία έχουμε για τις 10 μέρες καθώς όσο αυξάνονται οι μέρες τόσο μειώνεται ο συντελεστής Spearman και για τους δύο περιβαλλοντικούς παράγοντες χωρίς να βοηθάει στη ανάλυση μας.

Correlations					
			ΚΡΟΥΣΜΑΤΑ	ΜΕΣΟΣ_ΟΡΟΣ_ΘΕΡΜΟΚΡΑΣΙΑΣ	ΜΕΣΗ_ΣΧΕΤΙΚΗ_ΥΓΡΑΣΙΑ
Spearman's rho	ΚΡΟΥΣΜΑΤΑ	Correlation Coefficient	1,000	,367**	,314**
		Sig. (2-tailed)	.	,000	,000
		N	153	153	153
	ΜΕΣΟΣ_ΟΡΟΣ_ΘΕΡΜΟΚΡΑΣΙΑΣ	Correlation Coefficient	,367**	1,000	,346**
		Sig. (2-tailed)	,000	.	,000
		N	153	153	153
	ΜΕΣΗ_ΣΧΕΤΙΚΗ_ΥΓΡΑΣΙΑ	Correlation Coefficient	,314**	,346**	1,000
		Sig. (2-tailed)	,000	,000	.
		N	153	153	153
**. Correlation is significant at the 0.01 level (2-tailed).					

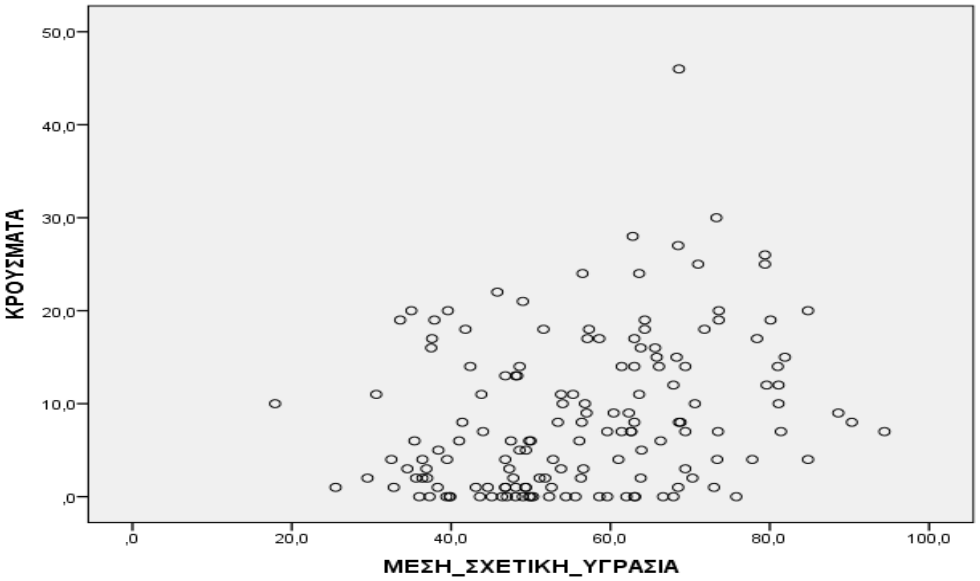
Πίνακας 4.2.Συσχέτιση 10ης ημέρας κρουσμάτων με μέση Υγρασία και μέση Θερμοκρασία

Σύμφωνα με τον πινάκα Correlations το p-value είναι αρκετά μικρότερο από το επίπεδο σημαντικότητας α=5% που σημαίνει ότι υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ κρουσμάτων και μέσης θερμοκρασίας καθώς και μεταξύ κρουσμάτων και μέσης υγρασίας. Οι μεταβλητές κρούσματα και θερμοκρασία είναι θετικά συσχετισμένες ωστόσο η (γραμμική) συσχέτιση των μεταβλητών μας είναι μέτρια καθώς ο συντελεστής του Spearman ισούται με 0,367 ενώ μεταξύ κρουσμάτων και υγρασίας ο συντελεστής του Spearman ισούται με 0,314 που σημαίνει ότι η συσχέτιση των μεταβλητών αυτών είναι μέτρια.



Διάγραμμα 4.1. Διασπορά κρουσμάτων και μέσης Θερμοκρασίας

Σύμφωνα με το διάγραμμα διασποράς διακρίνουμε θετική συσχέτιση μεταξύ των δύο ποσοτικών μεταβλητών χωρίς ωστόσο να είναι ισχυρή διότι οι τιμές είναι διασκορπισμένες χωρίς να έχουν συγκεκριμένη τάση ανόδου ή καθόδου.

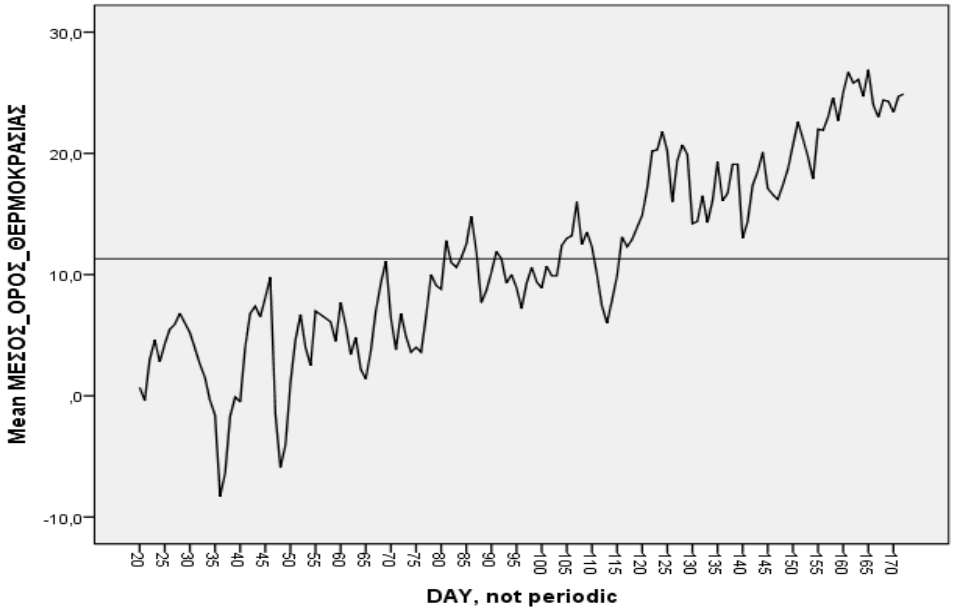


Διάγραμμα 4.2. Διασπορά κρουσμάτων και μέσης Υγρασίας

Στο διάγραμμα διασποράς μεταξύ κρουσμάτων και μέσης σχετικής υγρασίας παρατηρούμε μια θετική συσχέτιση μεταξύ τους ωστόσο η συσχέτιση αυτή δεν είναι ισχυρή καθώς οι περισσότερες τιμές συγκεντρώνονται στο διάστημα 38 έως 78 χωρίς να ακολουθούν κάποια συγκεκριμένη πορεία.

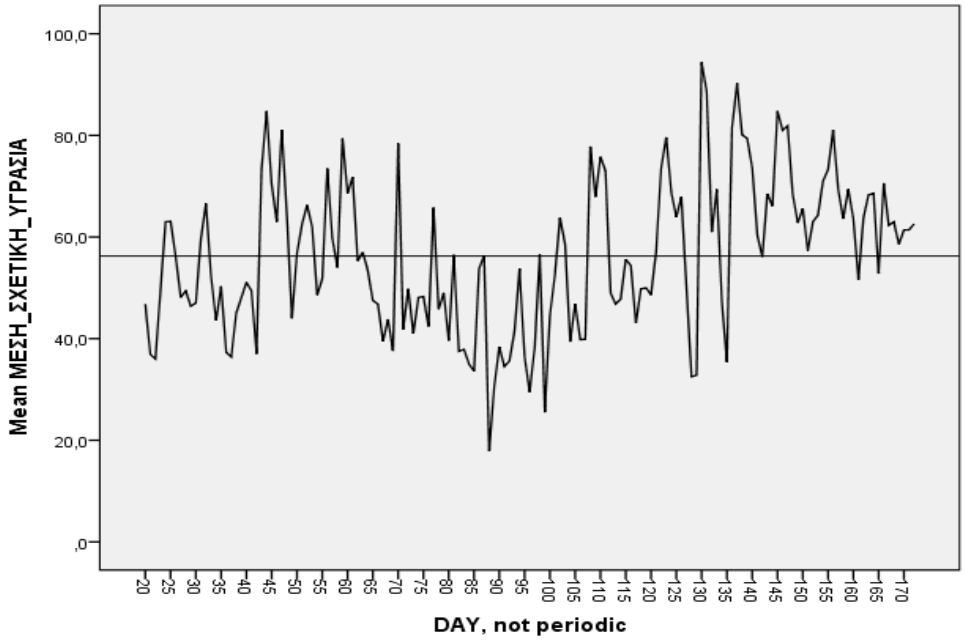
4.2 Διαγραμματική απεικόνιση των χρονοσειρών ‘‘Μέσος Όρος Θερμοκρασίας’’, ‘‘Μέση Σχετική Υγρασία’’ και ‘‘Κρούσματα’’

Στην ενότητα αυτή θα ασχοληθούμε διαγραμματικά για τις χρονοσειρές της μέσης θερμοκρασίας, της μέσης υγρασίας καθώς και των κρουσμάτων για μια πρώτη απεικόνιση της ανάλυσης μας.



Διάγραμμα 4.3.Χρονοσειρά μέσης Θερμοκρασίας

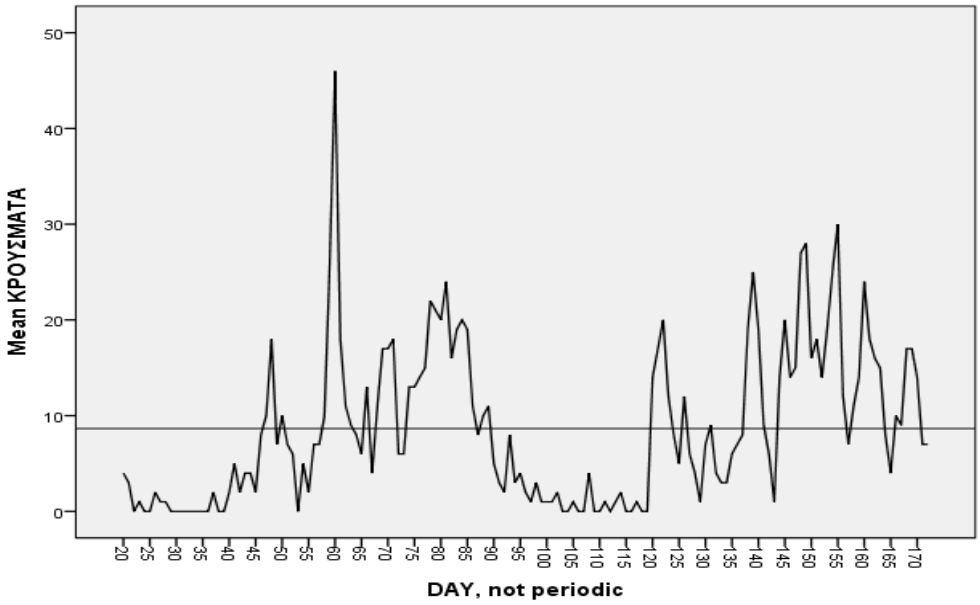
Σύμφωνα με το χρονόγραμμα παρατηρούμε ότι η μέση θερμοκρασία για την περίοδο 20/01/2020 μέχρι 20/6/2020 αρχικά είχε μια καθοδική πορεία αλλά με το πέρασμα του χρόνου αυξάνεται η θερμοκρασία, έχοντας τάση ανοδικής πορείας μέχρι το τέλος της χρονικής περιόδου καθώς οι τιμές της δεν κινούνται γύρω από το μέσο όρο της.



Διάγραμμα 4.4. Χρονοσειρά μέσης Υγρασίας

Αντιθέτως, όσον αφορά στο χρονόγραμμα για τη μέση σχετική υγρασία παρατηρείται μια πιο στάσιμη σειρά καθώς οι τιμές κυμαίνονται γύρω από το

μέσο όρο της χωρίς να υπάρχουν ουσιαστικές διαφοροποιήσεις στις διάφορες τιμές της για το συγκεκριμένο χρονικό έτος.



Διάγραμμα 4.5. Χρονοσειρά Κρουσμάτων

Από το παραπάνω χρονόγραμμα παρατηρούμε ότι έχουμε μια ακανόνιστη διαφόριση των κρουσμάτων στην Seoul διότι οι παρατηρήσεις απεικονίζονται ως ένα σύνολο μη κανονικών κινήσεων και ούτε μπορούμε να χαρακτηρίσουμε αυτή τη χρονολογική σειρά ως στάσιμη καθώς οι περισσότερες παρατηρήσεις δεν κυμαίνονται γύρω από το μέσο όρο της μεταβλητής μας.

4.2.1 Ανάλυση της χρονοσειράς ‘‘Κρούσματα’’

Στο σημείο αυτό θα χρησιμοποιήσουμε τη μέθοδο Box και Jenkins για την ανάλυση της χρονοσειράς μας μέσω των τριών σταδίων, δηλαδή την ταυτοποίηση της χρονοσειράς ως ARIMA (p,d,q) διαδικασία, την εκτίμηση των παραμέτρων p και q του υποδείγματος καθώς επίσης και τον έλεγχο του συγκεκριμένου υποδείγματος.

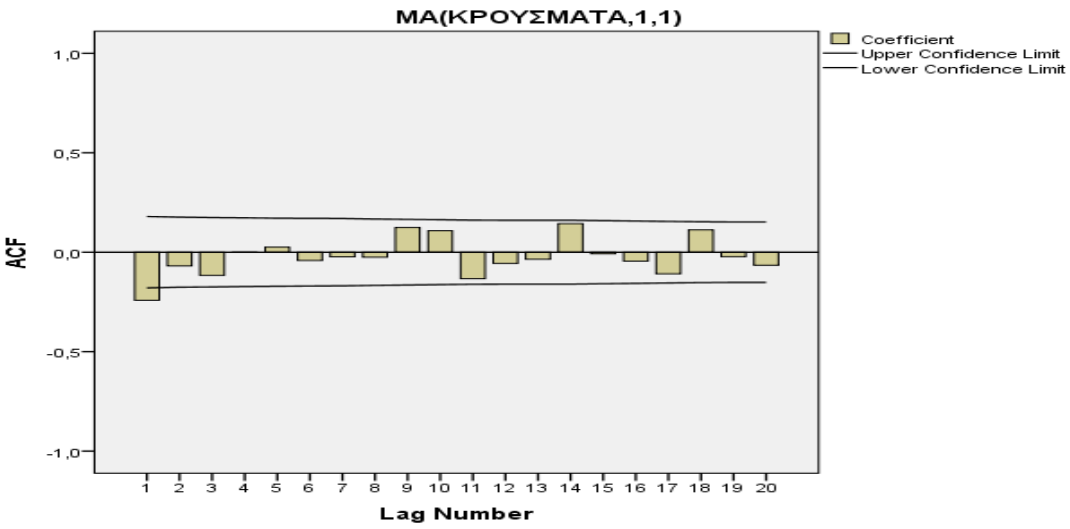
1) Ταυτοποίηση της χρονοσειράς

Σύμφωνα με τον συντελεστή συσχέτισης Spearman όσον αφορά τα κρούσματα θα χρησιμοποιήσουμε αυτά που αναφέρονται για 10 ημέρες καθώς έχουμε τη μεγαλύτερη συσχέτιση μεταξύ των κρουσμάτων και της μέσης θερμοκρασίας και της μέσης υγρασίας αντίστοιχα.

Autocorrelations					
Series: MA(ΚΡΟΥΣΜΑΤΑ,1,1)					
Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	-,242	,090	7,297	1	,007
2	-,069	,088	7,908	2	,019
3	-,116	,087	9,692	3	,021
4	-,001	,086	9,692	4	,046
5	,025	,085	9,781	5	,082
6	-,041	,085	10,019	6	,124
7	-,023	,085	10,093	7	,183
8	-,025	,083	10,185	8	,252
9	,124	,082	12,435	9	,190
10	,108	,082	14,187	10	,165
11	-,133	,081	16,922	11	,110
12	-,057	,080	17,425	12	,134
13	-,035	,080	17,619	13	,173
14	,144	,080	20,834	14	,106
15	-,008	,079	20,845	15	,142
16	-,045	,078	21,176	16	,172
17	-,108	,077	23,137	17	,145
18	,112	,076	25,283	18	,117
19	-,022	,076	25,369	19	,149
20	-,065	,076	26,106	20	,162
a. The underlying process assumed is independence (white noise).					
b. Based on the asymptotic chi-square approximation.					

Πίνακας 4.3.Αυτοσυσχετίσεις Κρουσμάτων

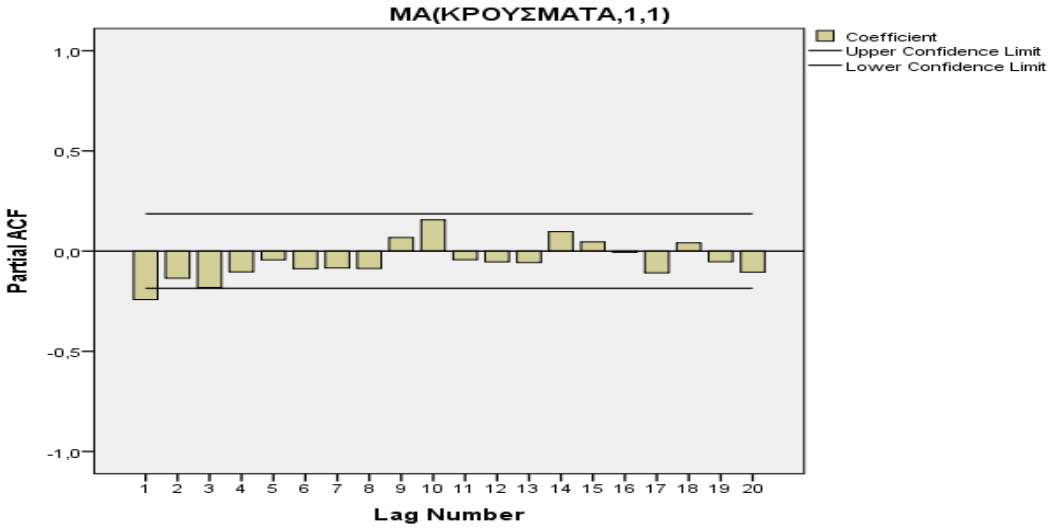
Σύμφωνα με τον πίνακα Autocorrelations και εφαρμόζοντας κεντρικό κινητό μέσο στα κρούσματα παρατηρούμε ότι οι αυτοσυσχετίσεις φθίνουν με γρήγορο ρυθμό και συγκλίνουν στο μηδέν και για αυτό θεωρείται ως στάσιμη χρονοσειρά ο αριθμός των κρουσμάτων. Επιπλέον από τις τιμές της στατιστικής Box-Ljung, τα αντίστοιχα p-values τους είναι αρκετά μεγαλύτερα από το επίπεδο σημαντικότητας α=1% , κάτι που δεν ισχύει για επίπεδο σημαντικότητας α=5% και για αυτό λόγο θα χρησιμοποιήσουμε α=1% όπου ο αριθμός των κρουσμάτων αποτελούν λευκό θόρυβο και συνεπώς δεν συσχετίζονται μεταξύ τους.



Διάγραμμα 4.6.Συντελεστές και διαστήματα εμπιστοσύνης

Partial Autocorrelations		
Series: MA(ΚΡΟΥΣΜΑΤΑ,1,1)		
Lag	Partial Autocorrelation	Std. Error
1	-,242	,093
2	-,135	,093
3	-,182	,093
4	-,104	,093
5	-,043	,093
6	-,088	,093
7	-,084	,093
8	-,086	,093
9	,068	,093
10	,156	,093
11	-,043	,093
12	-,054	,093
13	-,057	,093
14	,097	,093
15	,046	,093
16	-,006	,093
17	-,109	,093
18	,041	,093
19	-,053	,093
20	-,105	,093

Πίνακας 4.4.Μερικές Αυτοσυσχετίσεις Κρουσμάτων



Διάγραμμα 4.7.Συντελεστές και διαστήματα εμπιστοσύνης

Στο διάγραμμα μερικών αυτοσυσχετίσεων (PACF) απεικονίζονται οι συντελεστές των μερικών αυτοσυσχετίσεων καθώς και τα 95% διαστήματα εμπιστοσύνης τα οποία μας βοηθούν στην επιλογή των σημαντικών αυτοσυσχετίσεων. Στο παραπάνω διάγραμμα παρατηρούμε ότι μόνο μία αυτοσυσχέτιση είναι σημαντική την οποία θα χρησιμοποιήσουμε στη συνέχεια της ανάλυσής μας.

Σύμφωνα με τη μέθοδο Box και Jenkins ο αριθμός των κρουσμάτων είναι σημαντική μεταβλητή και για αυτό το λόγο θα προχωρήσουμε στα επόμενα στάδια, δηλαδή στην εκτίμηση των παραμέτρων p και q της ARIMA(p,d,q) διαδικασίας καθώς και στον έλεγχο σημαντικότητας του υποδείγματος.

2) Εκτίμηση των παραμέτρων p και q του υποδείγματος

Σύμφωνα με το πρώτο στάδιο παρατηρείται ότι υπάρχει μια μη μηδενική αυτοσυσχέτιση καθώς και μια μη μηδενική μερική αυτοσυσχέτιση οπότε θεωρούμε την χρονοσειρά ως μια διαδικασία κινητού μέσου πρώτης τάξης, δηλαδή ARIMA(0,1,1).

Model Statistics						
Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Normalized BIC	Statistics	DF	Sig.	
MA(ΚΡΟΥΣΜΑΤΑ,1,1)-Model_1	0	3,545	35,903	17	,005	0

Πίνακας 4.5.Στατιστική συνάρτηση του μοντέλου ARIMA

Στον πίνακα Model Statistics αναφέρεται η τιμή του κριτηρίου Normalized BIC που ισούται με 3,545, η τιμή της στατιστικής Q των Box και Ljung για 17 χρονικές

υστερήσεις που ισούται με 35,903 καθώς επίσης και το p-value=0,005 που είναι μικρότερο από το επίπεδο σημαντικότητας α=1% που σημαίνει τη μη ύπαρξη λευκού θορύβου στο συγκεκριμένο υπόδειγμα.

ARIMA Model Parameters								
					Estimate	SE	t	Sig.
MA(ΚΡΟΥΣΜΑΤΑ,1,1)-Model_1	MA(ΚΡΟΥΣΜΑΤΑ,1,1)	No Transformation	Constant		,021	,427	,048	,961
			Difference		1			
			MA	Lag 1	,077	,081	,940	,349

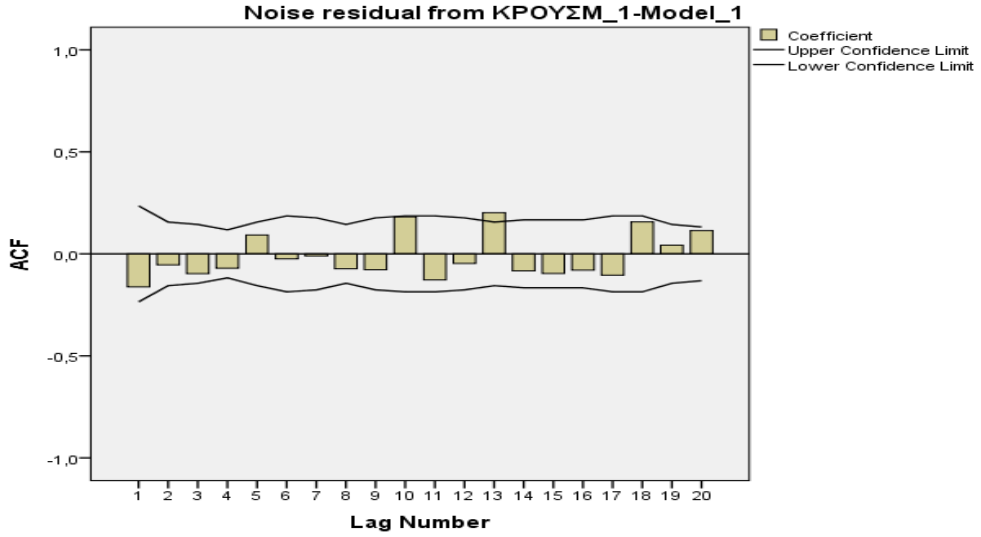
Πίνακας 4.6. Παράμετροι του μοντέλου ARIMA

Σύμφωνα με τον πίνακα ARIMA Model Parameters εκτιμάται η παράμετρος του κινητού μέσου πρώτης τάξης που ισούται με 0,077 και το p-value του ελέγχου ισούται με 0,349 και συνεπώς η παράμετρος είναι στατιστικά σημαντική για το υπόδειγμα.

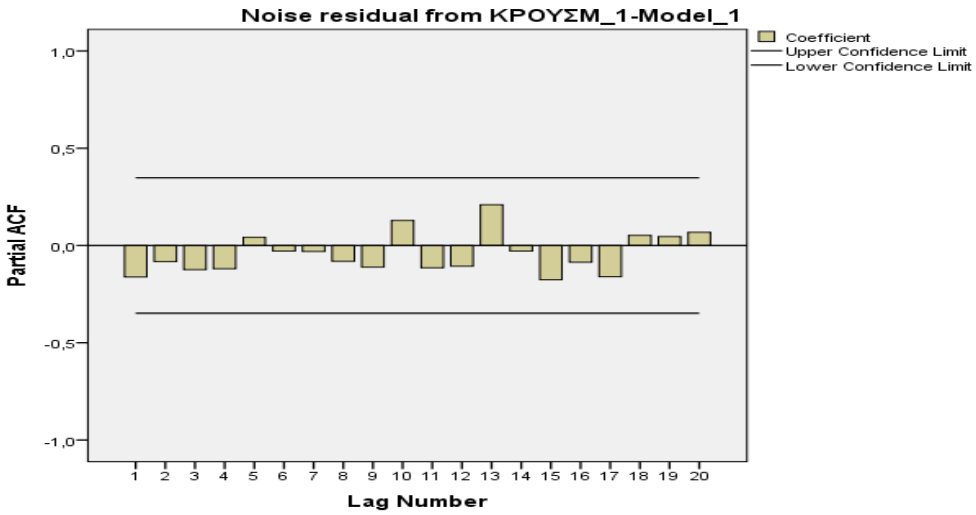
3)Έλεγχος του υποδείγματος

Στο τρίτο στάδιο θα πρέπει να ελέγξουμε κατά πόσο ικανοποιητικό είναι το εκτιμηθέν υπόδειγμα. Αυτό γίνεται ως εξής:

- Οι αυτοσυσχετίσεις και μερικές αυτοσυσχετίσεις της χρονοσειράς των σφαλμάτων να μην διαφέρουν σημαντικά από το μηδέν καθώς και ότι μια ή δύο αυτοσυσχετίσεις μπορούν να υπερβαίνουν το 95% διάστημα εμπιστοσύνης αλλά η διαφορά αυτή θα πρέπει να είναι αρκετά μικρή.
- Τα σφάλματα θα πρέπει να αποτελούν λευκό θόρυβο



Διάγραμμα 4.8. Συντελεστές και διαστήματα εμπιστοσύνης



Διάγραμμα 4.9. Συντελεστές και διαστήματα εμπιστοσύνης

Autocorrelations					
Series: Noise residual from KPOYΣM_1-Model_1					
Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	-,162	,118	1,897	1	,168
2	-,054	,078	2,381	2	,304
3	-,097	,072	4,193	3	,241
4	-,070	,059	5,607	4	,230
5	,092	,078	7,011	5	,220
6	-,024	,093	7,079	6	,314
7	-,010	,088	7,093	7	,419
8	-,073	,072	8,114	8	,422
9	-,078	,088	8,890	9	,447
10	,182	,093	12,701	10	,241
11	-,127	,093	14,578	11	,203
12	-,047	,088	14,862	12	,249
13	,202	,078	21,592	13	,062
14	-,083	,083	22,584	14	,067
15	-,096	,083	23,917	15	,067
16	-,080	,083	24,835	16	,073
17	-,105	,093	26,101	17	,073
18	,157	,093	28,930	18	,049
19	,042	,072	29,274	19	,062
20	,114	,066	32,299	20	,040
a. The underlying process assumed is independence (white noise).					
b. Based on the asymptotic chi-square approximation.					

Πίνακας 4.7. Αυτοσυσχετίσεις Κρουσμάτων

Από τα διαγράμματα των αυτοσυσχετίσεων και μερικών αυτοσυσχετίσεων παρατηρούμε ότι μόνο μια αυτοσυσχέτιση υπερβαίνει το 95% διάστημα εμπιστοσύνης χωρίς όμως να απέχει αρκετά. Επίσης, τα p-value είναι αρκετά μεγαλύτερα από το επίπεδο σημαντικότητας α=1% και συνεπώς τα σφάλματα αποτελούν λευκό θόρυβο. Όλα τα παραπάνω υποδεικνύουν ότι το υπόδειγμά μας είναι ικανοποιητικό και μπορεί να χρησιμοποιηθεί για περαιτέρω ανάλυση.

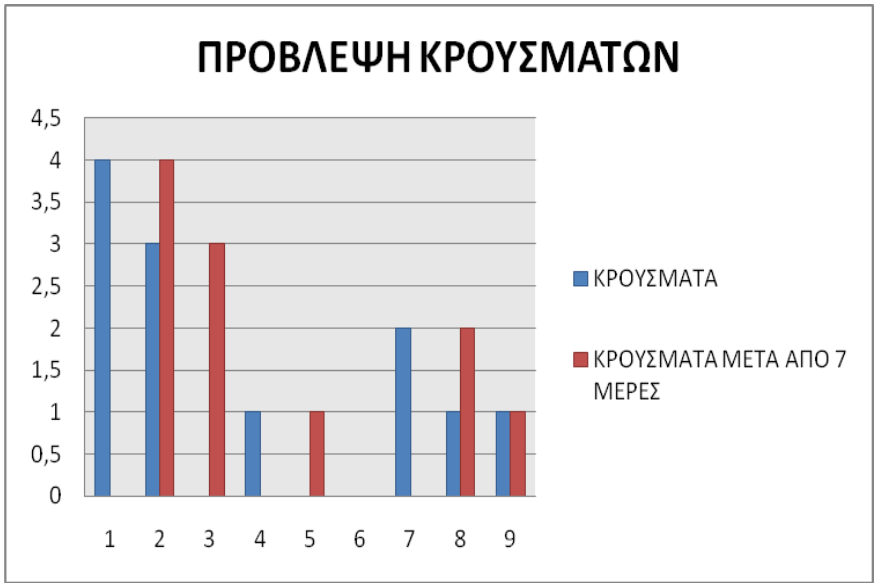
4.2.2 Πρόβλεψη των μελλοντικών τιμών της χρονοσειράς ‘‘Κρούσματα’’

Σε αυτή την ενότητα θα ασχοληθούμε με την πρόβλεψη μελλοντικών τιμών του αριθμού των κρουσμάτων μετά από επτά μέρες. Ενδεικτικά παρουσιάζουμε τον παρακάτω πίνακα.

ΚΡΟΥΣΜΑΤΑ	ΚΡΟΥΣΜΑΤΑ ΜΕΤΑ ΑΠΟ 7 ΜΕΡΕΣ
4	0
3	4
0	3
1	0
0	1
0	0
2	0
1	2
1	1

Πίνακας 4.8.Μελλοντικές τιμές Κρουσμάτων

Στον παρακάτω πίνακα οι τιμές μπορεί να είναι οι ίδιες καθώς το υπόδειγμα ARIMA(0,1,1) που χρησιμοποιήσαμε υπολογίζει προβλέψεις για μια μόνο μελλοντική περίοδο αλλά οι προβλέψεις αυτές ξεκινάνε για 21/1/2020 ενώ τα αρχικά μας κρούσματα ξεκινάνε για 20/1/2020.



Διάγραμμα 4.10.Μελλοντική πρόβλεψη Κρουσμάτων

ΣΥΝΟΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Όπως έχουμε αναφέρει ο κύριος στόχος της παρούσας διπλωματικής εργασίας είναι να προσδιορίσουμε ποιοι δημογραφικοί και περιβαλλοντικοί παράγοντες συμβάλλουν στην εξάπλωση του κορωνοϊού. Παρατηρήσαμε ότι ο κορωνοϊός είχε ταχεία διάδοση μέσα στο 2020 ξεπερνώντας τα επίπεδα που έφταναν παλιότερα σημαντικές ασθένειες όπως η γρίπη, η πνευμονία και το κρυολόγημα όπως επίσης ότι τα περισσότερα κρούσματα εμφανίστηκαν είτε σε νοσοκομεία, είτε σε ιατρικά κέντρα και γενικότερα σε μέρη που αφορά την ιατρική περίθαλψη.

Χρησιμοποιώντας την λογιστική παλινδρόμηση στην ανάλυσή μας συμπεραίναμε ότι το φύλο, η ηλικία και η επαρχία από την οποία προέρχεται ο κάθε πολίτης της Νότιας Κορέας είχε σημαντικό ρόλο στην εξάπλωση της συγκεκριμένης ασθένειας καθώς και ότι οι περισσότεροι ασθενείς επιβίωναν. Όσον αφορά τους περιβαλλοντικούς παράγοντες καταλήξαμε στο συμπέρασμα ότι η μέση θερμοκρασία και η μέση υγρασία σχετίζονται με την αύξηση των κρουσμάτων ενώ οι υπόλοιποι παράγοντες δεν συμβάλλουν ιδιαίτερα στην εξάπλωση του COVID-19. Τέλος, εφαρμόζοντας την μεθοδολογία Box and Jenkins για την ανάλυση χρονολογικών σειρών παρατηρήσαμε ότι για την επαρχία Seoul ο αριθμός των κρουσμάτων δεν αυξάνεται σταδιακά αλλά παραμένει ο ίδιος.

ΠΑΡΑΡΤΗΜΑ

- **Δεδομένα της παρούσας εργασίας**

Στο παρακάτω παράρτημα σας παραθέτουμε αναλυτικά την επεξήγηση των αρχείων τα οποία χρησιμοποιήσαμε για την παραπάνω ανάλυση.

Το αρχείο Case.txt έχει τις εξής στήλες:

Case_id	Κωδικό για κάθε περίπτωση μόλυνσης
Province	Στην επαρχία που εμφανίστηκε η μόλυνση
City	Στην πόλη όπου εμφανίζεται η μόλυνση
Group	Στο γκρουπ στο οποίο ανήκει κάθε περίπτωση: αληθής, δηλαδή ότι μολύνθηκε και ψευδής, δηλαδή ότι δεν μολύνθηκε
Infection_case	Συγκεκριμένο μέρος που εμφανίζεται η μόλυνση
Confirmed	Αριθμός επιβεβαιωμένων κρουσμάτων
Latitude	Ακριβή τοποθεσία μόλυνσης με βάση το γεωγραφικό πλάτος
Longitude	Ακριβή τοποθεσία μόλυνσης με βάση το γεωγραφικό μήκος

Το αρχείο PatientInfo.txt έχει τις εξής στήλες:

Patient_id	Κωδικός ασθενή
Sex	Φύλο ασθενή, άντρας ή γυναίκα
Age	Η ηλικία του κάθε ασθενή
Country	Η χώρα από την οποία προέρχεται ο ασθενής
Province	Η επαρχία από την οποία προέρχεται ο ασθενής
City	Η πόλη από την οποία προέρχεται ο ασθενής
Infection_case	Συγκεκριμένο μέρος που εμφανίζεται η ασθένεια
Infected_by	Ο κωδικός ασθενή που κόλλησε άλλον ασθενή
Contact_number	Ο αριθμός των ατόμων που ήρθε σε επαφή ο ασθενής

Symptom_onset_date	Ημερομηνία εμφάνισης των συμπτωμάτων

Confirmed_date	Ημερομηνία επιβεβαίωσης της ασθένειας
Released_date	Ημερομηνία που έφυγε από το νοσοκομείο
Deceased_date	Ημερομηνία που απεβίωσε ο ασθενής
State	Κατάσταση: απομόνωσης/απελευθέρωσης/απεβίωση

Το αρχείο Policy.txt έχει τις εξής στήλες:

Policy_id	Κωδικός πολιτικής
Country	Η χώρα στην οποία εφαρμόστηκε η πολιτική
Type	Το είδος της πολιτικής που χωρίζεται σε μετανάστευση, εκπαίδευση, υγεία, τεχνολογία, κοινωνία και άλλα
Gov_policy	Η πολιτική που εφαρμόζει η κυβέρνηση
Detail	Οι λεπτομέρειες της πολιτικής
Start_date	Ημερομηνία έναρξης της πολικής
End_date	Ημερομηνία λήξης της πολιτικής

Το αρχείο Region.txt έχει τις εξής στήλες:

Code	Κωδικός περιοχής
Province	Η επαρχία από την οποία προέρχεται ο ασθενής
City	Η πόλη από την οποία προέρχεται ο ασθενής
Latitude	Ακριβή τοποθεσία με βάση το γεωγραφικό πλάτος
Longitude	Ακριβή τοποθεσία με βάση το γεωγραφικό μήκος
Elementary_school_count	Αριθμός δημοτικών σχολείων
Kindergarten_count	Αριθμός νηπιαγωγείων
University_count	Αριθμός πανεπιστημίων

Academy_ratio	Η αναλογία των ακαδημιών
Elderly_population_ratio	Η αναλογία του πληθυσμού των ηλικιωμένων

Elderly_alone_ratio	Η αναλογία των ηλικιωμένων που μένουν μόνοι τους
Nursing_home_count	Ο αριθμός των γηροκομείων

Το αρχείο SearchTrend.txt έχει τις εξής στήλες:

Date	Ημερομηνία
Cold	Το ποσοστό κρυολογημάτων στην Κορέα
Flu	Το ποσοστό γρίπης στην Κορέα
Pneumonia	Το ποσοστό πνευμονίας στην Κορέα
Coronavirus	Το ποσοστό κορωνοϊού στην Κορέα

Το αρχείο Time.txt έχει τις εξής στήλες:

Date	Ημερομηνία
Time	Χρόνος όπου το 0 αναφέρεται στις 12:00 πμ. ενώ το 16 αναφέρεται στις 04:00 μμ.
Test	Ο συσσωρευμένος αριθμός των τεστ
Negative	Ο συσσωρευμένος αριθμός των αρνητικών τεστ
Confirmed	Ο συσσωρευμένος αριθμός των θετικών τεστ
Released	Ο συσσωρευμένος αριθμός απελευθερώσεων
Deceased	Ο συσσωρευμένος αριθμός αποθανόντων

Το αρχείο TimeAge.txt έχει τις εξής στήλες:

Date	Ημερομηνία

Time	Χρόνος
Age	Η ηλικία των ασθενών
Confirmed	Ο συσσωρευμένος αριθμός των θετικών τεστ
Deceased	Ο συσσωρευμένος αριθμός αποθανόντων

Το αρχείο TimeGender.txt έχει τις εξής στήλες:

Date	Ημερομηνία
Time	Χρόνος
Sex	Φύλο ασθενή, άντρας ή γυναίκα
Confirmed	Ο συσσωρευμένος αριθμός των θετικών τεστ
Deceased	Ο συσσωρευμένος αριθμός αποθανόντων

Το αρχείο TimeProvince.txt έχει τις εξής στήλες:

Date	Ημερομηνία
Time	Χρόνος
Province	Η επαρχία από την οποία προέρχεται ο ασθενής
Confirmed	Ο συσσωρευμένος αριθμός των επιβεβαιωμένων στην επαρχία
Released	Ο συσσωρευμένος αριθμός των απελευθερώσεων στην επαρχία
Deceased	Ο συσσωρευμένος αριθμός των αποθανόντων στην επαρχία

Το αρχείο Weather.txt έχει τις εξής στήλες:

Code	Κωδικός της κάθε περιοχής
Province	Πόλη/Επαρχία
Date	Ημερομηνία
Avg_temp	Μέσος όρος θερμοκρασίας

Min_temp	Μικρότερη θερμοκρασία
Max_temp	Μεγαλύτερη θερμοκρασία
Precipitation	Η καθημερινή βροχόπτωση
Max_wind_speed	Η μέγιστη ταχύτητα ανέμου
Most_wind_direction	Η πιο συχνή κατεύθυνση ανέμου
Avg_relative_humidity	Η μέση σχετική υγρασία

- **Πίνακες συχνοτήτων για τον τόπο κατοικίας και το μέρος εμφάνισης του κορωνοϊού για το 2ο κεφάλαιο**

Στο παρακάτω παράρτημα δίνονται οι αναλυτικοί πίνακες συχνοτήτων για τον τόπο κατοικίας και το μέρος εμφάνισης του κορωνοϊού που χρησιμοποιήθηκαν στην ανάλυσή μας.

Ο πίνακας συχνοτήτων για τον τόπο κατοικίας είναι:

ΠΟΛΗ					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	94	1,8	1,8	1,8
	Andong-si	53	1,0	1,0	2,8
	Ansan-si	33	,6	,6	3,5
	Anseong-si	4	,1	,1	3,6
	Anyang-si	63	1,2	1,2	4,8
	Asan-si	18	,3	,3	5,1
	Bonghwa-gun	71	1,4	1,4	6,5
	Bucheon-si	162	3,1	3,1	9,6
	Buk-gu	28	,5	,5	10,2
	Bupyeong-gu	83	1,6	1,6	11,8
	Busanjin-gu	14	,3	,3	12,1
	Buyeo-gun	13	,3	,3	12,3
	Changnyeong-gun	9	,2	,2	12,5
	Changwon-si	31	,6	,6	13,1
	Cheonan-si	110	2,1	2,1	15,2
	Cheongdo-gun	43	,8	,8	16,1
	Cheongju-si	20	,4	,4	16,4
	Cheongsong-gun	2	,0	,0	16,5
	Cheorwon-gun	11	,2	,2	16,7
	Chilgok-gun	51	1,0	1,0	17,7
	Chuncheon-si	10	,2	,2	17,9
	Chungju-si	13	,3	,3	18,1
	Daedeok-gu	5	,1	,1	18,2
	Dalseo-gu	29	,6	,6	18,8
	Dalseong-gun	1	,0	,0	18,8
	Dalsung-gun	6	,1	,1	18,9
	Danyang-gun	1	,0	,0	18,9
	Dobong-gu	62	1,2	1,2	20,1
	Dong-gu	40	,8	,8	20,9

	Dongdaemun-gu	38	,7	,7	21,6
	Dongducheon-si	5	,1	,1	21,7
	Dongjak-gu	53	1,0	1,0	22,8
	Dongnae-gu	39	,8	,8	23,5
	etc	110	2,1	2,1	25,7
	Eumseong-gun	7	,1	,1	25,8
	Eunpyeong-gu	53	1,0	1,0	26,8
	Gangbuk-gu	24	,5	,5	27,3
	Gangdong-gu	35	,7	,7	28,0
	Ganghwa-gun	1	,0	,0	28,0
	Gangnam-gu	83	1,6	1,6	29,6
	Gangneung-si	9	,2	,2	29,8
	Gangseo-gu	83	1,6	1,6	31,4
	Gapyeong-gun	1	,0	,0	31,4
	Geochang-gun	19	,4	,4	31,8
	Geoje-si	13	,3	,3	32,0
	Geumcheon-gu	32	,6	,6	32,6
	Geumjeong-gu	7	,1	,1	32,8
	Gijang-gun	1	,0	,0	32,8
	Gimcheon-si	24	,5	,5	33,2
	Gimhae-si	14	,3	,3	33,5
	Gimje-si	3	,1	,1	33,6
	Gimpo-si	37	,7	,7	34,3
	Goesan-gun	11	,2	,2	34,5
	Gongju-si	3	,1	,1	34,6
	Goryeong-gun	10	,2	,2	34,8
	Goseong-gun	2	,0	,0	34,8
	Goyang-si	53	1,0	1,0	35,8
	Gumi-si	78	1,5	1,5	37,3
	Gunpo-si	75	1,5	1,5	38,8
	Gunsan-si	4	,1	,1	38,9
	Gunwi-gun	6	,1	,1	39,0
	Guri	1	,0	,0	39,0
	Guri-si	11	,2	,2	39,2
	Guro-gu	85	1,6	1,6	40,9
	Gwacheon-si	9	,2	,2	41,0
	Gwanak-gu	113	2,2	2,2	43,2
	Gwangjin-gu	17	,3	,3	43,5
	Gwangju-si	35	,7	,7	44,2
	Gwangmyeong-si	28	,5	,5	44,8
	Gwangyang-si	1	,0	,0	44,8

	Gyeongju-si	52	1,0	1,0	45,8
	Gyeongsan-si	639	12,4	12,4	58,2
	Gyeryong-si	3	,1	,1	58,2
	Gyeyang-gu	55	1,1	1,1	59,3
	Haeundae-gu	22	,4	,4	59,7
	Haman-gun	1	,0	,0	59,7
	Hamyang-gun	1	,0	,0	59,7
	Hanam-si	12	,2	,2	60,0
	Hapcheon-gun	10	,2	,2	60,2
	Hoengseong-gun	1	,0	,0	60,2
	Hongseong-gun	4	,1	,1	60,3
	Hwaseong-si	41	,8	,8	61,1
	Hwasun-gun	1	,0	,0	61,1
	Icheon-dong	1	,0	,0	61,1
	Icheon-si	15	,3	,3	61,4
	Iksan-si	2	,0	,0	61,4
	Inje-gun	1	,0	,0	61,5
	Jangsu-gun	1	,0	,0	61,5
	Jeju-do	18	,3	,3	61,8
	Jeonju-si	8	,2	,2	62,0
	Jeungpyeong-gun	2	,0	,0	62,0
	Jincheon-gun	1	,0	,0	62,0
	Jinju-si	12	,2	,2	62,3
	Jongno-gu	22	,4	,4	62,7
	Jung-gu	55	1,1	1,1	63,8
	Jungnang-gu	33	,6	,6	64,4
	Kyeongsan-si	1	,0	,0	64,4
	Mapo-gu	38	,7	,7	65,2
	Michuhol-gu	69	1,3	1,3	66,5
	Miryang-si	5	,1	,1	66,6
	Mokpo-si	5	,1	,1	66,7
	Muan-gun	2	,0	,0	66,7
	Mungyeong-si	4	,1	,1	66,8
	Nam-gu	36	,7	,7	67,5
	Namdong-gu	45	,9	,9	68,4
	Namhae-gun	1	,0	,0	68,4
	Namyangju-si	47	,9	,9	69,3
	Nonsan-si	6	,1	,1	69,4
	Nowon-gu	43	,8	,8	70,2
	Okcheon-gun	1	,0	,0	70,3
	Osan-si	8	,2	,2	70,4

	Paju-si	16	,3	,3	70,7
	Pocheon-si	19	,4	,4	71,1
	Pohang-si	53	1,0	1,0	72,1
	Pyeongtaek-si	61	1,2	1,2	73,3
	Sacheon-si	3	,1	,1	73,4
	Saha-gu	11	,2	,2	73,6
	Samcheok-si	1	,0	,0	73,6
	Sancheong-gun	1	,0	,0	73,6
	Sangju-si	15	,3	,3	73,9
	sankyeock-dong	1	,0	,0	73,9
	Sasang-gu	5	,1	,1	74,0
	Sejong	52	1,0	1,0	75,0
	Seo-gu	78	1,5	1,5	76,5
	Seocheon-gun	1	,0	,0	76,6
	Seocho-gu	55	1,1	1,1	77,6
	Seodaemun-gu	36	,7	,7	78,3
	Seongbuk-gu	33	,6	,6	79,0
	Seongdong-gu	49	,9	,9	79,9
	Seongju-gun	23	,4	,4	80,3
	Seongnam-si	173	3,3	3,3	83,7
	Seosan-si	9	,2	,2	83,9
	Siheung-si	27	,5	,5	84,4
	Sokcho-si	3	,1	,1	84,5
	Songpa-gu	55	1,1	1,1	85,5
	Suncheon-si	2	,0	,0	85,6
	Suseong-gu	14	,3	,3	85,8
	Suwon	1	,0	,0	85,8
	Suwon-si	100	1,9	1,9	87,8
	Suyeong-gu	11	,2	,2	88,0
	Taean-gun	1	,0	,0	88,0
	Taebaek-si	1	,0	,0	88,0
	Uijeongbu-si	47	,9	,9	88,9
	Uiseong-gun	41	,8	,8	89,7
	Uiwang-si	10	,2	,2	89,9
	Ulju-gun	6	,1	,1	90,0
	Wonju-si	25	,5	,5	90,5
	Yangcheon-gu	71	1,4	1,4	91,9
	Yangju-si	9	,2	,2	92,1
	Yangpyeong-si	1	,0	,0	92,1
	Yangsan-si	8	,2	,2	92,3
	Yecheon-gun	44	,9	,9	93,1

	Yeongcheon-si	38	,7	,7	93,8
	Yeongdeok-gun	2	,0	,0	93,9
	Yeongdeungpo-gu	62	1,2	1,2	95,1
	Yeongju-si	5	,1	,1	95,2
	Yeongwol-gun	1	,0	,0	95,2
	Yeongyang-gun	2	,0	,0	95,2
	Yeonje-gu	4	,1	,1	95,3
	Yeonsu-gu	45	,9	,9	96,2
	Yeosu-si	3	,1	,1	96,2
	Yongin-si	104	2,0	2,0	98,3
	Yongsan-gu	50	1,0	1,0	99,2
	Yuseong-gu	40	,8	,8	100,0
	Total	5165	100,0	100,0	

- Μέρη εμφάνισης του κορωνοϊού:

ΠΕΡΙΠΤΩΣΕΙΣΜΟΛΥΝΣΗΣ					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	919	17,8	17,8	17,8
	Anyang Gunpo Pastors Group	1	,0	,0	17,8
	Biblical Language study meeti	3	,1	,1	17,9
	Bonghwa Pureun Nursing Home	31	,6	,6	18,5
	Changnyeong Coin Karaoke	4	,1	,1	18,5
	Cheongdo Daenam Hospital	21	,4	,4	19,0
	contact with patient	1610	31,2	31,2	50,1
	Coupang Logistics Center	80	1,5	1,5	51,7
	Daejeon door-to-door sales	1	,0	,0	51,7
	Daezayeon Korea	3	,1	,1	51,8
	Day Care Center	43	,8	,8	52,6
	Dongan Church	17	,3	,3	52,9
	Dunsan Electronics Town	13	,3	,3	53,2
	etc	703	13,6	13,6	66,8
	Eunpyeong St. Mary's Hospital	16	,3	,3	67,1
	Gangnam Dongin Church	1	,0	,0	67,1
	Gangnam Yeoksam-dong gatherin	6	,1	,1	67,2
	Geochang Church	6	,1	,1	67,3
	Geumcheon-gu rice milling mac	6	,1	,1	67,5
	Guri Collective Infection	5	,1	,1	67,6
	Guro-gu Call Center	112	2,2	2,2	69,7
	Gyeongsan Cham Joeun Communit	10	,2	,2	69,9
	Gyeongsan Jeil Silver Town	12	,2	,2	70,1
	Gyeongsan Seorin Nursing Home	15	,3	,3	70,4
	gym facility in Cheonan	30	,6	,6	71,0
	gym facility in Sejong	4	,1	,1	71,1
	Itaewon Clubs	162	3,1	3,1	74,2
	KB Life Insurance	13	,3	,3	74,5

	Korea Campus Crusade of Chris	7	,1	,1	74,6
	Milal Shelter	11	,2	,2	74,8
	Ministry of Oceans and Fisher	28	,5	,5	75,4
	Onchun Church	33	,6	,6	76,0
	Orange Life	1	,0	,0	76,0
	Orange Town	7	,1	,1	76,2
	overseas inflow	840	16,3	16,3	92,4
	Pilgrimage to Israel	2	,0	,0	92,5
	Richway	128	2,5	2,5	94,9
	River of Grace Community Chur	1	,0	,0	95,0
	Samsung Fire & Marine Insuran	4	,1	,1	95,0
	Samsung Medical Center	7	,1	,1	95,2
	Seocho Family	5	,1	,1	95,3
	Seongdong-gu APT	13	,3	,3	95,5
	Seoul City Hall Station safet	3	,1	,1	95,6
	Shincheonji Church	107	2,1	2,1	97,7
	SMR Newly Planted Churches Gr	36	,7	,7	98,4
	Suyeong-gu Kindergarten	3	,1	,1	98,4
	Uiwang Logistics Center	2	,0	,0	98,5
	Wangsung Church	24	,5	,5	98,9
	Yangcheon Table Tennis Club	44	,9	,9	99,8
	Yeonana News Class	5	,1	,1	99,9
	Yeongdeungpo Learning Institu	3	,1	,1	99,9
	Yongin Brothers	4	,1	,1	100,0
	Total	5165	100,0	100,0	

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

1. Ευαγγελάρας Χ. (2012), *Ανάλυση δεδομένων με τη χρήση στατιστικών πακέτων:* Σημειώσεις για το SPSSV19, Πανεπιστήμιο Πειραιώς, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πρόγραμμα Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

2. Αντζουλάκος Δ. (2018), *Ανάλυση Επιβίωσης*, Πανεπιστήμιο Πειραιώς, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πρόγραμμα Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

3. Σαχλάς Αθ. και Μπερσίμης Σ. (2014-2015), *Βιοστατιστική και στατιστικές μέθοδοι στην Επιδημιολογία*, Πανεπιστήμιο Πειραιώς, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πρόγραμμα Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

4. Φράγκος Χρήστος Κων.(2004), *Μεθοδολογία Έρευνας Αγοράς και Ανάλυση Δεδομένων με τη χρήση του Στατιστικού Πακέτου SPSS FOR WINDOWS*, Εκδοτικός Οίκος Interbooks, Ιπποκράτους 18 Αθήνα

Διαδικτυακές Ιστοσελίδες

1. Forbes, 12/11/2020, Η μυστική ιστορία του πρώτου κορωνοϊού: https://www.capital.gr/forbes/3445509/i-mustiki-istoria-tou-protou-koronoiou

2. Υγεία, Μάρτιος 2020,Covid-19: Πανδημία 2020: https://www.hygeia.gr/koronoios-ti-prepei-na-gnorizoyme-amp-pos-mporoyme-na-profylachthoyme/

3. Centers for Disease Control and Prevention, 7/06/2021, Similarities and Differences between Flu and Covid-19: https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm

4. Centers for Disease Control and Prevention: https://www.cdc.gov/

5. Worldometers: https://www.worldometers.info/

6. Our in world in data, 5/03/2021,Emerging Covid-19 success story: South Korea learned the lessons of MERS: https://ourworldindata.org/covid-exemplar-south-korea

7. Korea Disease Control and Prevention Agency: http://www.kdca.go.kr/cdc_eng/

8. https://www.kaggle.com/

9. https://eclass.aegean.gr/modules/document/file.php/SAS142/%CE%A3%CE%B7%CE%BC%CE%B5%CE%B9%CF%89%CF%83%CE%B5%CE%B9%CF%82%20SPSS%2017%20Forecasting.pdf

10. https://eclass.aegean.gr/modules/document/file.php/SAS142/KEF6.pdf