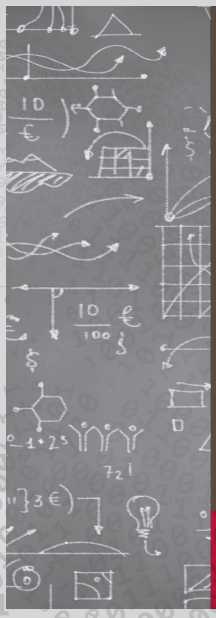
ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ (Ε.Σ.Ι.)
GREEK STATISTICAL INSTITUTE (G.S.I.)

30ο ΠΑΝΕΛΛΗΝΙΟ ΣΥΝΕΔΡΙΟ ΣΤΑΤΙΣΤΙΚΗΣ



ΑΝΑΛΥΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ: *Εφαρμογές σε Τουρισμό, Υγεία, Περιβάλλον*

Πρακτικά 30ου Πανελλήνιου Συνεδρίου Στατιστικής

Proceedings of the 30th Panhellenic Statistics Conference

Λάρνακα, 20-22 Απριλίου 2017

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
(Ε.Σ.Ι)
GREEK STATISTICAL INSTITUTE
(G.S.I)

Π Ρ Α Κ Τ Ι Κ Α

30ου Πανελληνίου
Συνεδρίου Στατιστικής

P R O C E E D I N G S

of the 30th Panhellenic
Statistics Conference

Ανάλυση Μεγάλων Δεδομένων: Εφαρμογές σε Τουρισμό, Υγεία, Περιβάλλον

Big Data: Applications in Tourism, Healthcare and Environment

Πύλα Λάρνακας, 20-22 Απριλίου 2017

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
(Ε.Σ.Ι)
GREEK STATISTICAL INSTITUTE
(G.S.I.)

Π Ρ Α Κ Τ Ι Κ Α

30ου Πανελληνίου
Συνεδρίου Στατιστικής

Ανάλυση Μεγάλων Δεδομένων: Εφαρμογές σε Τουρισμό, Υγεία, Περιβάλλον

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΠΑΝΕΠΙΣΤΗΜΙΟ UCLan CYPRUS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΕΧΝΟΛΟΓΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

Πύλα Λάρνακας,, 20-22 Απριλίου 2017

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
Σολωμού 5 (Πλατεία Εξαρχείων)
Τηλ.& Fax 210 33.03.909
Email: esi-stat@hol.gr
http://www.esi-stat.gr

ISBN: 978-618-83805-0-9
ISSN: 1792-2461

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Το 30ο Πανελλήνιο Συνέδριο Στατιστικής διοργανώθηκε από το Ελληνικό Στατιστικό Ινστιτούτο (ΕΣΙ) στην Πύλα την περίοδο 20-22 Απριλίου 2017, σε συνεργασία με το Πανεπιστήμιο UCLan Cyprus, το Πανεπιστήμιο Κύπρου και το Τεχνολογικό Πανεπιστήμιο Κύπρου. Το θέμα του Συνεδρίου ήταν *Ανάλυση Μεγάλων Δεδομένων: Εφαρμογές σε Τουρισμό, Υγεία, Περιβάλλον.*

Έλαβαν μέρος περίπου πενήντα (50) σύνεδροι από την Ελλάδα και την Κύπρο. Στο Συνέδριο προσκεκλημένοι ομιλητές ήταν η *Ruth Heller*, Associate Professor, Department of Statistics and Operations Research at Tel-Aviv University, Israel, ο *Finn Lindgren*, Professor, School of Mathematics, University of Edinburgh, ο *Αθανάσιος Θανόπουλος*, Πρόεδρος της Ελληνικής Στατιστικής Αρχής, καθώς και ο *Σωτήριος Μπερσίμης*, Επίκουρος Καθηγητής Πανεπιστημίου Πειραιώς και Πρόεδρος του ΕΟΠΥΥ.

Το Επιστημονικό Πρόγραμμα περιελάμβανε σαράντα πέντε (45) ανακοινώσεις σε παράλληλες συνεδρίες σε θέματα Στατιστικής, Πρόβλεψης-Χρονοσειρών, Στατιστικής στην Υγεία και την Τεχνολογία, Μπεϋζιανής Στατιστικής-Προσομοίωσης, Εφαρμοσμένης Στατιστικής, Πιθανοτήτων, Στοχαστικών Διαδικασιών κ.ά. Στο πλαίσιο του 30ου Πανελληνίου Συνεδρίου Στατιστικής πραγματοποιήθηκε και ειδική συνεδρία με θέμα *Κοινωνική και Εκπαιδευτική Στατιστική.*

Η εναρκτήρια τελετή του Συνεδρίου πραγματοποιήθηκε στο Πανεπιστήμιο UCLan Cyprus. Την έναρξη του Συνεδρίου κήρυξε ο Πρόεδρος του ΕΣΙ, τέως Αναπληρωτής Καθηγητής Χαράλαμπος Δαμιανού. Χαιρετισμούς απηύθυναν η Πρόεδρος της Οργανωτικής Επιτροπής, Λέκτορας Μιλτώ Χατζηκυριάκου και ο Αντιπρύτανης του Πανεπιστημίου UCLan Cyprus, Καθηγητής Πανίκος Πουτζιουρής. Η τελετή έναρξης συνεχίστηκε με την απονομή του Ελένειου Βραβείου Καλύτερης Διδακτορικής Διατριβής στη Στατιστική 2015-2016 από τον Πρόεδρο του ΕΣΙ, Χαράλαμπο Δαμιανού. Το βραβείο απονεμήθηκε από κοινού στις κα. Μαρία Πίτσιλλου για τη διδακτορική διατριβή της με τίτλο «Testing Serial Dependence by the Distance Covariance Function» και κα. Ηλιάνα Χρίστου για την διδακτορική διατριβή της με τίτλο «A non-iterative method for fitting the single index quantile regression model with uncensored and censored data». Μέλη της επιτροπής του Ελένειου Βραβείου ήταν ο Καθηγητής Κωνσταντίνος Ζωγράφος, ο Καθηγητής Σταύρος Κουρούκλης και ο Αν. Καθηγητής Νικόλαος Παπαδάτος. Η κα. Μαρία Πίτσιλλου εκπόνησε τη διδακτορική της διατριβή στο Πανεπιστήμιο Κύπρου με τον Καθηγητή Κωνσταντίνο Φωκιανό,

ενώ η Ηλιάνα Χρίστου εκπόνησε τη διδακτορική διατριβή της στο Pennsylvania State Univeristy με τον Καθηγητή Michael Akritas.

Για το βραβείο είχε υποβληθεί και η διατριβή του κ. Κωνσταντίνου Α. Τασιά με τίτλο: «Development and analysis of advanced adaptive statistical process control charts for the joint monitoring of variables' central tendency and dispersion», University of Western Macedonia με επιβλέποντα (supervisor) τον Καθηγητή George Nenes.

Οι κοινωνικές εκδηλώσεις του συνεδρίου ήταν πλούσιες και ποικίλες. Την Πέμπτη το απόγευμα πραγματοποιήθηκε ξενάγηση και Welcome Cocktail στο *Μουσείο Θάλασσας* στην Αγία Νάπα. Την Παρασκευή το απόγευμα πραγματοποιήθηκε για του συνέδρους και τα συνοδεύοντα μέλη εκδρομή στην πόλη της Λάρνακας κατά την οποία έκαναν βόλτα στις Φοινικούδες και ξεναγήθηκαν στον Ιερό Ναό του Αγίου Λαζάρου, στο Τέμενος Χαλά Σουλτάν και αλλού και το βράδυ ακολούθησε δείπνο στο *Lysia Restaurant,* στην Πύλα, στην περιοχή της Ορόκλινης.

Στον τόμο αυτό περιλαμβάνονται εργασίες που παρουσιάστηκαν στο Συνέδριο και υποβλήθηκαν για δημοσίευση. Όλες οι εργασίες κρίθηκαν από κριτές με την φροντίδα των υπευθύνων έκδοσης πρακτικών.

Οι παρατηρήσεις και τα σχόλια των κριτών, σύμφωνα με την πάγια πολιτική που ακολουθεί το ΕΣΙ, αφορούν κυρίως στον τρόπο παρουσίασης της εργασίας και στην παρουσία ή όχι τυπογραφικών και σοβαρών επιστημονικών λαθών. Οι εργασίες πρέπει να έχουν ικανό στατιστικό περιεχόμενο, να αναδεικνύουν το πρόβλημα που μελετούν, να μην περιορίζονται μόνο σε Περιγραφική Στατιστική. Για το σκοπό αυτό υπάρχουν κριτήρια δημοσίευσης εργασιών στα Πρακτικά του ΕΣΙ τα οποία είναι αναρτημένα στην ιστοσελίδα του ΕΣΙ, www.esi-stat.gr. Όλες οι εργασίες, για τις οποίες ζητήθηκε αναθεώρηση, κρίθηκαν εκ νέου από τους κριτές ή από τους υπεύθυνους έκδοσης των πρακτικών.

Συνολικά υποβλήθηκαν δώδεκα (12) εργασίες, από τις οποίες μία εργασία ανακλήθηκε από τους συγγραφείς και δύο απορρίφθηκαν, σύμφωνα με απόφαση του Δ.Σ. και ύστερα από σχετική πρόταση των κριτών και των υπεύθυνων της έκδοσης των Πρακτικών. Ως κριτές των εργασιών συνεργάστηκαν οι: Καλαματιανού Αγλαΐα, Μοσχονά Θεανώ-Εριφύλλη, Μπατσίδης Απόστολος, Οικονόμου Πολυχρόνης, Παπαϊωάννου Θοδωρής, Παπαϊωάννου Τάκης, Πετρόπουλος Κωνσταντίνος, Πολυχρονόπουλος Ευάγγελος, Σοφιανοπούλου Χρύσα, Τζαβελάς Γεώργιος, Φουσκάκης Δημήτριος. Η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις

ευχαριστίες της προς όλους τους κριτές για την επιμελημένη και προσεκτική αξιολόγηση των εργασιών.

Η σειρά παρουσίασης των εργασιών στον παρόντα τόμο είναι αλφαβητική με βάση το επώνυμο του πρώτου συγγραφέα.

Το Διοικητικό Συμβούλιο του ΕΣΙ αισθάνεται την ανάγκη να ευχαριστήσει το Πανεπιστήμιο UCLan Cyrpus, το Πανεπιστήμιο Κύπρου, το Τεχνολογικό Πανεπιστήμιο Κύπρου και την Οργανωτική Επιτροπή για την πολύ καλή οργάνωση και προσφορά τους.

Τέλος, το ΔΣ του ΕΣΙ εκφράζει τις ευχαριστίες του στη Γραμματέα του ΕΣΙ, Μαρία Χιώλου, για την τεχνική επιμέλεια της έκδοσης των Πρακτικών και των CD.

ΕΚ ΜΕΡΟΥΣ ΤΟΥ ΔΣ ΤΟΥ ΕΣΙ

Οι υπεύθυνοι Έκδοσης Πρακτικών 30^{ου} Συνεδρίου

Χαράλαμπος Δαμιανού	Σόνια Μαλεφάκη
Δημοσθένης Παναγιωτάκος	Τάκης Παπαϊωάννου

ΧΟΡΗΓΟΙ ΣΥΝΕΔΡΙΟΥ

ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ

Πέμπτη, 20 Απριλίου 2017

16:00-17:00	**Εγγραφή Συνέδρων - Διανομή συνεδριακού υλικού**
ΑΜΦΙΘΕΑΤΡΟ 2 **17:00- 17:30**	**Έναρξη του Συνεδρίου - Χαιρετισμοί**

ΑΜΦΙΘΕΑΤΡΟ 2	**ΟΜΙΛΙΕΣ ΣΕ ΟΛΟΜΕΛΕΙΑ** **ΕΛΕΝΕΙΟ ΒΡΑΒΕΙΟ ΚΑΛΥΤΕΡΗΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ 2015-2016** **Προεδρεύων: Χ. Δαμιανού**
17:30	**Pitsillou M.:** Testing serial dependence by the distance covariance function **E. Christou**: A non-iterative method for fitting the single index quantile regression model with uncensored and censored data
18:30	**ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΟΜΙΛΙΑ** **Προεδρεύουσα: Μπαξεβάνη Α.**
	F. Lindgren: Quantifying the uncertainty of contour maps

19:30	**Welcome Cocktail** **Μουσείο Θάλασσας (*Αγία Νάπα*)**

Παρασκευή, 21 Απριλίου 2017

ΑΜΦΙΘΕΑΤΡΟ 2	ΟΜΙΛΙΑ ΣΕ ΟΛΟΜΕΛΕΙΑ ΠΡΟΣΚΕΚΛΗΜΕΝΗ ΟΜΙΛΙΑ **Προεδρεύων: κ. Φωκιανός**
09:30	**R. Heller:** Multivariate tests of associations based on univariate tests

ΑΙΘΟΥΣΑ Α	ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ **Προεδρεύων: Κούτρας Μ.**
10:30	**M. Hadjikyriakou:** Asymptotic results for demimartingales and for randomly indexed demimartingales
10:50	**A. Baxevani:** Use of latent transformed Gaussian processes for certain environmental processes
11:10	**S. Vakeroudis:** Windings of planar stochastic processes and applications
11:30	**Ν. Μαχαιράς, Σ. Τζανίνης:** Αλλαγή μέτρου για σύνθετες ανανεωτικές διαδικασίες με εφαρμογές στις αρχές υπολογισμού ασφαλίστρου

ΑΙΘΟΥΣΑ Β	**ΣΤΑΤΙΣΤΙΚΗ** **Προεδρεύουσα: Μπαξεβάνη Α.**
10:30	**P. Doukhan, K. Fokianos, B. Stove, D. Tjostheim:** Modeling and inference for multivariate count time series
10:50	**F. Panagou, D. Karlis:** Use of Gaussian copula for clustering mixed mode data
11:10	**L. Smallman, A. Artemiou, J. Morgan:** Sparse generalised PCA with applications to text data
11:30	**Δ. Πανάρετος, Μ. Βαμβακάρη, Γ. Τζαβελάς, Δ. Παναγιωτάκος:** Διερεύνηση του ρόλου των ορθογωνίων μετασχηματισμών των αξόνων στην παραγοντική ανάλυση σε σχέση με την επαναληψιμότητα των εξαγόμενων προτύπων κάτω από διάφορα σενάρια τυχαίου σφάλματος στα αρχικά δεδομένα

11:50	***ΔΙΑΛΕΙΜΜΑ-ΚΑΦΕΣ***

ΑΙΘΟΥΣΑ Α	**ΣΤΑΤΙΣΤΙΚΗ** **Προεδρεύων: Δαμιανού Χ**
12:20	**E. Manoli, T. Christofides:** An improved version of the item count technique and protection of privacy
12:40	**Ι. Παπατσούμα, Ν. Φαρμάκης:** Εκτίμηση του συντελεστή μεταβλητότητας από δεδομένα διακριτής ομοιόμορφης κατανομής
13:00	**Μ. Ταφιάδη, Γ. Ηλιόπουλος:** Ακριβείς έλεγχοι για τον λόγο των παραμέτρων κλίμακας δύο κατανομών Laplace

ΑΙΘΟΥΣΑ Β	**ΠΡΟΒΛΕΨΗ- ΧΡΟΝΟΣΕΙΡΕΣ** **Προεδρεύων: Φωκιανός Κ.**
12:20	**D. Chorozoglou, D. Kugiumtzis, E. Papadimitriou:** Testing the small-worldness property of earthquake networks: case study on time series from seismicity in Greece
12:40	**A. Milionis, N. Galanopoulos:** Time series with interdependent level and second moment: Testing, consequences for modelling, applications
13:00	**H. Tsangari:** Modeling and projecting heat-related mortality

13:30-14:30	**ΓΕΥΜΑ ΣΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ** ***χορηγία του Πανεπιστημίου UCLan Cyprus***

ΑΙΘΟΥΣΑ Α	**ΠΙΘΑΝΟΤΗΤΕΣ** **Προεδρεύουσα: Χατζηκυριάκου Μ.**
15:00	**Ch. Charalambous, T. Christofides:** Distance between a U-statistic and a normal random variable
15:20	**A. N. Arapis, F. S. Makri, Z. M. Psillakis:** Joint distribution of run statistics in zero-one Markov dependent trials
15:40	**D. Christofides, K. Markström:** Matrix-valued random variables and applications
16:00	**Μ. Β. Κούτρας, Π. Μπομποτάς:** Κατανομή του ελάχιστου από τυχαίο αριθμό τυχαίων μεταβλητών

ΑΙΘΟΥΣΑ Β	**ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ** **Προεδρεύων: Δαμιανού Χ.**
15:00	**V. Chasiotis, S. Kounias, N. Farmakis:** Optimal equally replicated designs of multi-level factors minimizing the E(f_{NOD}) criterion
15:20	**Θ. Χατζηπαντελής, Μ. Σωτήρογλου:** Η διδακτική και μαθησιακή προσέγγιση των Ποσοτικών Μεθόδων στις Κοινωνικές Επιστήμες
15:40	**S. Kitromilidou:** Estimation of research activity based on patent data
16:00	**N. Thanasoulias, D. Tsorova:** Use of business intelligence tools for the evaluation of drinking water quality in water distribution networks

17:00-19:00	***ΕΚΔΡΟΜΗ*** ***Ξενάγηση στην πόλη της Λάρνακας***

21:00	**ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ** ***Lysia Restaurant***

Σάββατο, 22 Απριλίου 2017

ΑΜΦΙΘΕΑΤΡΟ 2	ΟΜΙΛΙΕΣ ΣΕ ΟΛΟΜΕΛΕΙΑ **ΠΡΟΣΚΕΚΛΗΜΕΝΕΣ ΟΜΙΛΙΕΣ** **Προεδρεύων: Χριστοφίδης Τ.**
09:30	**Α. Θανόπουλος :** Big Data in Official Statistics with applications in Tourism, Healthcare and Environment
10:30	**Σ. Μπερσίμης:** Η επιστήμη της Στατιστικής ως εργαλείο για τον έλεγχο των δαπανών υγείας

11:30	***ΔΙΑΛΕΙΜΜΑ-ΚΑΦΕΣ***

ΑΙΘΟΥΣΑ Α	**ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΥΓΕΙΑ** **Προεδρεύων: Φαρμάκης Ν.**
12:00	**Μ. Χουζούρης, J. Hyosung, Π. Ξένος, Α. Μπαλασοπούλου, Γ. Χαραλάμπους:** Μέτρηση της πιθανότητας εισαγωγής ασθενών σε μονάδες νοσηλείας μέσω των ΤΕΠ
12:20	**Σ. Μπερσίμης, Α. Σαχλάς, R. Sparks:** Παρακολούθηση και αξιολόγηση της απόδοσης στις Υπηρεσίες Υγείας
12:40	**Β. Ισμυρλής:** Παράγοντες που επηρεάζουν την υποκειμενική υγεία των ευρωπαίων πολιτών: Μια στατιστική ανάλυση της ευρωπαϊκής κοινωνικής έρευνας
13:00	**Π. Καράκος, Θ. Λιαλιάρης, Α. Καράκος:** Μελέτη επίδρασης του βιοσυντονισμού στον ανθρώπινο οργανισμό

ΑΙΘΟΥΣΑ Β	**ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑ** **Προεδρεύουσα: Τσαγκάρη Χ.**
12:00	**Ι. Τσιμπερίδης, Α. Καράκος:** Έλεγχος ημερήσιας χρήσης του υπολογιστή στα πλαίσια εγκληματολογικής έρευνας
12:20	**I. Lamprianou:** Crowd-sourced Twitter analysis: Issues of reliability and validity
12:40	**A. Santourian, G. Farmakis:** Access to administrative data sources for official statistics
13:00	**Κ. Παπά:** Conjoint analysis - Εφαρμογή στα κινητά τηλέφωνα smartphones

13:30-14:30	***ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ***

ΑΙΘΟΥΣΑ Α	**ΕΙΔΙΚΗ ΣΥΝΕΔΡΙΑ** **ΚΟΙΝΩΝΙΚΗ ΚΑΙ ΕΚΠΑΙΔΕΥΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ** **Προεδρεύων: Μιχαηλίδης Μ.**
14:40	**L. Kyriakides, Ch. Charalambous, I. Televantou:** The contribution of schooling to the cognitive development of education students in Cyprus: An application of regression discontinuity approach in educational effectiveness research
15:00	**Μ. Π. Μιχαηλίδης:** Χρήση λανθανόντων μοντέλων κατά τη διερεύνηση της παραγοντικής δομής ψυχομετρικών εργαλείων
15:20	**Γ. Σπανούδης, Α. Τούρβα:** Μελετώντας τη σχέση του χρόνου οπτικής επιθεώρησης με τη νοημοσύνη στα παιδιά

15:40	**Ε. Παπαναστασίου:** Αναπάντητες ερωτήσεις σε εξεταστικά δοκίμια και η σημασία τους στην αξιολόγηση

ΑΙΘΟΥΣΑ Β	**ΜΠΕΫΖΙΑΝΗ ΣΤΑΤΙΣΤΙΚΗ-ΠΡΟΣΟΜΟΙΩΣΗ** **Προεδρεύων: Μωυσιάδης Π.**
14:40	**S. Agapiou, D. Sanz-Alonso, O. Papaspiliopoulos, A. Stuart:** The intrinsic dimension of importance sampling
15:00	**K. Mantzouni, I. Ntzoufras, M. Kateri:** Bayesian testing for association models in contingency tables using power priors
15:20	**Ζ. Ζωσιμάς, Α. Μπουρνέτας:** Μελέτη βελτιώσεων προσομοίωσης μεθόδου τυχαίων δέντρων για αποτίμηση δικαιωμάτων αμερικάνικου τύπου
15:40	**Π. Μωυσιάδης:** ΤΖΑΚΠΟΤ και Μαθηματικά

ΑΝΑΡΤΗΜΕΝΗ ΑΝΑΚΟΙΝΩΣΗ	**Χ. Χαραλάμπους, Στ.Α. Χατζόπουλος, Φ. Κολυβά-Μαχαίρα, Λ. Αγγελής, Θ. Τσιτσώνη, Σ. Παπαδοπούλου, Ζ. Ανδρεοπούλου:** Ακαδημαϊκές επιδόσεις στο Α.Π.Θ.: Έμφυλη διάσταση, κοινωνική ευαλωτότητα

ΑΜΦΙΘΕΑΤΡΟ 2 **16:00**	**ΟΛΟΜΕΛΕΙΑ: ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ**

ΚΟΙΝΩΝΙΚΕΣ ΕΚΔΗΛΩΣΕΙΣ

Πέμπτη 20 Απριλίου 2017

Welcome Cocktail: Μουσείο Θάλασσας (*Αγία Νάπα*)
Ώρα: 19:30

Παρασκευή 21 Απριλίου 2017

Εκδρομή: 17:00-19:00

Επίσημο Δείπνο Συνεδρίου
Προορισμός: *Lysia Restaurant*
Ώρα: 21:00

Επιστημονική Επιτροπή

***Δαμιανού Χαράλαμπος**, Πανεπιστήμιο Αθηνών*
***Δονάτος Γεώργιος**, Πανεπιστήμιο Αθηνών*
***Ζωγράφος Κωνσταντίνος**, Πανεπιστήμιο Ιωαννίνων*
***Κάκουλλος Θεόφιλος,** Πανεπιστήμιο Αθηνών*
***Καραγρηγορίου Αλέξανδρος**, Πανεπιστήμιο Αιγαίου*
***Καρλής Δημήτριος**, Οικονομικό Πανεπιστήμιο Αθηνών*
***Κουνιάς Στρατής**, Πανεπιστήμιο Αθηνών*
***Κουτρουβέλης Ιωάννης**, Πανεπιστήμιο Πατρών*
***Κυριακούσης Ανδρέας**, Χαροκόπειο Πανεπιστήμιο*
***Μπερσίμης Σωτήριος**, Πανεπιστήμιο Πειραιώς*
***Μπουρνέτας Απόστολος**, Πανεπιστήμιο Αθηνών*
***Μωυσιάδης Πολυχρόνης**, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
***Παναγιωτάκος Δημοσθένης**, Χαροκόπειο Πανεπιστήμιο*
***Παπαϊωάννου Τάκης**, Πανεπιστήμια Πειραιώς και Ιωαννίνων*
***Πετρόπουλος Κωνσταντίνος**, Πανεπιστήμιο Πατρών*
***Τσακλίδης Γεώργιος**, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*
***Φουσκάκης Δημήτριος**, Μετσόβειο Πολυτεχνείο*
***Φωκιανός Κωνσταντίνος**, Πανεπιστήμιο Κύπρου*
***Χατζηκυριάκου Μιλτώ**,UCLan Cyprus*
***Χριστοφή Κώστας**, Τεχνολογικό Πανεπιστήμιο Κύπρου*
***Χριστοφίδης Τάσος,** Πανεπιστήμιο Κύπρου*

Οργανωτική Επιτροπή

***Μαυρικίου Πετρούλα**, Πανεπιστήμιο Frederick*
***Μίτλεττον Νίκος**, Τεχνολογικό Πανεπιστήμιο Κύπρου*
***Μπαξεβάνη Αναστασία**, Πανεπιστήμιο Κύπρου*
***Φωκιανός Κωνσταντίνος**, Πανεπιστήμιο Κύπρου*
***Χατζηκυριάκου Μιλτώ,** UCLan Cyprus*
***Χριστοφή Κώστας**, Τεχνολογικό Πανεπιστήμιο Κύπρου*
***Χριστοφίδης Δημήτρης**, UCLan Cyprus*
***Χριστοφίδης Τάσος**, Πανεπιστήμιο Κύπρου*

Επιτροπή Προγράμματος
Γεώργιος Ηλιόπουλος

εργασίες

στα ελληνικά

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 30ου Πανελληνίου Συνεδρίου Στατιστικής (2017), σελ.20-34

ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΗΝ ΥΠΟΚΕΙΜΕΝΙΚΗ ΥΓΕΙΑ ΤΩΝ ΕΥΡΩΠΑΙΩΝ ΠΟΛΙΤΩΝ: ΜΙΑ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΤΗΣ ΕΥΡΩΠΑΪΚΗΣ ΚΟΙΝΩΝΙΚΗΣ ΕΡΕΥΝΑΣ

Ισμυρλής Βασίλειος
Πανεπιστήμιο Μακεδονίας
vasismir@gmail.com

ΠΕΡΙΛΗΨΗ

Η υγεία θεωρείται ως ένας καθοριστικός παράγοντας που μπορεί να επηρεάζει σημαντικά την ευημερία των πολιτών, την οποία οι κυβερνήσεις των κρατών αντιμετωπίζουν με μεγάλη σοβαρότητα και ανησυχία. Ένα σημαντικό ζήτημα είναι ποιοι παράγοντες επηρεάζουν την κατάσταση της υγείας των πολιτών. Σε αυτήν την έρευνα αναζητούνται κυρίως κοινωνικοί παράγοντες, οι οποίοι μπορούν να αποδειχθούν κρίσιμοι για τον καθορισμό της κατάστασης της υποκειμενικής αντίληψης για την υγεία. Για αυτό το λόγο αναλύονται δεδομένα από μια μεγάλη κοινωνική έρευνα που διεξάγεται στην Ευρώπη κάθε δύο χρόνια. Η έρευνα αυτή είναι η Ευρωπαϊκή Κοινωνική Έρευνα (European Social Survey), η οποία παρέχει αρκετά σημαντικά στοιχεία για την κοινωνική, οικονομική και πολιτισμική ζωή των ευρωπαίων πολιτών. Από αυτήν την έρευνα αναζητήθηκαν οι πιο κρίσιμοι κοινωνικοί, οικονομικοί και πολιτισμικοί παράγοντες που επηρεάζουν την υποκειμενική αντίληψη της υγείας. Για να επιτευχθεί αυτό, χρησιμοποιήθηκε μια στατιστική μεθοδολογία, η Παραγοντική ανάλυση των αντιστοιχιών, η οποία έχει τη δυνατότητα να αναδεικνύει με εύκολο τρόπο τις διαφοροποιήσεις μεταξύ των μεταβλητών ή/και των τιμών τους. Τα αποτελέσματα ανέδειξαν ως καθοριστικότερους παράγοντες που επηρεάζουν την υγεία: το εισόδημα, την κοινωνικότητα, την κατάθλιψη, τη γενικότερη ικανοποίηση και ευτυχία. Επίσης βρέθηκε ότι κάποιοι άλλοι παράγοντες δεν επηρεάζουν την υγεία, όπως η οικογενειακή κατάσταση, η θρησκεία, η κατοικία κ.α.

Λέξεις Κλειδιά: Ευρωπαϊκή κοινωνική έρευνα, Ευημερία, Υγεία, Υποκειμενική αντίληψη υγείας, Παραγοντική ανάλυση των αντιστοιχιών, Κοινωνικοί/οικονομικοί/πολιτισμικοί παράγοντες.

1. ΕΙΣΑΓΩΓΗ

Η υγεία θεωρείται ένας σημαντικός παράγοντας που επηρεάζει την καθημερινότητα του ανθρώπου και θεωρείται καθοριστικός για την αξιολόγηση της ευημερίας του. Οι

κυβερνώντες των χωρών έχοντας καταλάβει τη σημαντικότητα της υγείας, προσπαθούν να εξασφαλίσουν την καλύτερη δυνατή για τους πολίτες, παρόλο που λόγω της παγκόσμιας οικονομικής κρίσης, υπάρχουν πολλά εμπόδια για να το πετύχουν. Έτσι έρευνες που εξετάζουν την κατάσταση της υγείας των πολιτών θεωρούνται απαραίτητες και γίνονται συχνά είτε από τα ίδια τα κράτη ή άλλους οργανισμούς και φορείς.

Εκτός από την αντικειμενική κατάσταση της υγείας, η οποία είναι μετρήσιμη με καθαρά ιατρικούς όρους, υπάρχει και η υποκειμενική αντίληψη για την υγεία. Η αντίληψη αυτή διαμορφώνεται λαμβάνοντας υπόψη και άλλους παράγοντες και καταλήγει στην εκτίμηση ενός ατόμου για την κατάσταση της υγείας του. Το ζήτημα που πρέπει να ερευνηθεί, είναι ποιοι παράγοντες μπορούν να επηρεάσουν την αντίληψη αυτή. Οι παράγοντες αυτοί όπως έχει προκύψει από άλλες έρευνες, είναι κυρίως οικονομικοί, κοινωνικοί και πολιτισμικοί. Στο πεδίο των επιστημών υγείας, η μέτρηση υποκειμενικών μεγεθών δεν χρησιμοποιείται ευρέως και έχει δοθεί μεγαλύτερη έμφαση σε μετρήσιμα μεγέθη (Fotso & Kuate-Defo, 2005). Αυτό συμβαίνει παρόλο ότι υπάρχουν αποδείξεις, ότι η προσωπική αντίληψη για την υγεία μπορεί να παρέχει ενδείξεις για τις δυνατότητες του συστήματος υγείας, τη θνησιμότητα και άλλες σχετικές με την υγεία μεταβλητές (Gataulinas & Bancevita, 2014), που μπορεί να είναι εξαιρετικά χρήσιμες.

Για να αναζητηθούν δεδομένα που είναι σχετικά με τέτοιους παράγοντες επιλέχθηκε μια έρευνα κοινωνικής φύσεως που διεξάγεται στην Ευρώπη κάθε δύο έτη. Είναι η Ευρωπαϊκή Κοινωνική Έρευνα (ΕΚΕ), η οποία διεξάγεται δειγματοληπτικά με συνεντεύξεις σε ευρωπαίους πολίτες και περιέχει ερωτήσεις για πολλά κοινωνικά ζητήματα. Στην παρούσα εργασία χρησιμοποιήθηκαν δεδομένα από την έβδομη έκδοση αυτής της έρευνας, η οποία διεξήχθη το 2014.

Για την ανάλυση των δεδομένων της παραπάνω έρευνας, χρησιμοποιήθηκε μια μέθοδος από τον κλάδο της πολυδιάστατης στατιστικής ανάλυσης, η Παραγοντική Ανάλυση Αντιστοιχιών (ΠΑΑ). Η μέθοδος αυτή ενδείκνυται όταν τα δεδομένα είναι πολλών διαστάσεων και όταν δεν θεωρούμε ότι τα δεδομένα ακολουθούν μια θεωρητική κατανομή. Επιπλέον με τη δυνατότητα αναπαράστασης των αναλύσεων σε γραφικές παραστάσεις, δίνει μια ακόμη καλύτερη άποψη για την εξαγωγή συμπερασμάτων.

Η ΠΑΑ χρησιμοποιήθηκε στην ανάλυση των δεδομένων της ΕΚΕ. Έχοντας ως κύρια μεταβλητή την υποκειμενική αντίληψη για την υγεία, αναζητήθηκαν ποιοι παράγοντες την επηρεάζουν περισσότερο. Αποδείχθηκε ότι οι κυριότεροι παράγοντες που επηρεάζουν την υγεία είναι το εισόδημα, η κοινωνικότητα, η κατάθλιψη, η γενικότερη ικανοποίηση και η ευτυχία. Επίσης βρέθηκε ότι κάποιοι άλλοι παράγοντες δεν επηρεάζουν σημαντικά την υγεία, όπως η οικογενειακή κατάσταση, η θρησκεία, η κατοικία κ.α.

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Ευρωπαϊκή κοινωνική έρευνα (ΕΚΕ) – European Social Survey (ESS)

Η ΕΚΕ είναι μια έρευνα που διεξάγεται κάθε δύο χρόνια σε χώρες της Ε.Ε. με κύριο σκοπό την προσέγγιση των στάσεων των ευρωπαίων πολιτών σε κοινωνικά, πολιτικά και πολιτισμικά θέματα. Η έρευνα αυτή καθιερώθηκε το 2002 και η τελευταία πραγματοποιήθηκε το 2016. Στοιχεία δημοσιευμένα υπάρχουν ως και την 7^{η} έκδοση της έρευνας που πραγματοποιήθηκε το 2014 (ESS, 2014). Η Ελλάδα συμμετείχε σε αυτή τη μεγάλη έρευνα, δυστυχώς μόνο κατά τα έτη: 2002, 2004, 2008 και 2010.

Η έρευνα συμπεριλαμβάνει αρκετές ερωτήσεις που είναι κοινές σε όλες τις εκδόσεις της και άλλες που αλλάζουν κάθε δύο έτη. Στην έρευνα η μονάδα που ερευνείται είναι το άτομο, αλλά επιπλέον ρωτούνται και στοιχεία για άτομα που μένουν στο ίδιο νοικοκυριό. Επίσης συμπεριλαμβάνονται πολλές δημογραφικές μεταβλητές, όπως: φύλο, χώρα κατοικίας και καταγωγής, ηλικία, επάγγελμα, θρήσκευμα, πολιτικό κόμμα, επίπεδο εκπαίδευσης, επίπεδο εισοδήματος, τύπος εργασίας, μέρος κατοικίας (πόλη, χωριό), οικογενειακή κατάσταση, μόρφωση γονιών.

Η έρευνα αυτή συμπεριλαμβάνει τις παρακάτω θεματικές ενότητες:

- Πολιτική
- Υγεία
- Υποκειμενική ευημερία
- Απόψεις για μετανάστες
- Κοινωνική εμπιστοσύνη
- Κοινωνικοδημογραφικό προφίλ
- Κλίμακα ανθρώπινων αξιών

Για παράδειγμα η 7^{η} έκδοση της έρευνας για την οποία υπάρχουν διαθέσιμα δεδομένα στην ιστοσελίδα http://www.europeansocialsurvey.org/, πραγματοποιήθηκε σε 21 ευρωπαϊκές χώρες, είχε δείγμα 40185 ατόμων, το ποσοστό απόκρισης ήταν 70% και περιείχε 662 μεταβλητές-ερωτήσεις. Τα δεδομένα της έρευνας αυτής, θα αναλυθούν στην παρούσα εργασία.

2.2. Υγεία

Η υγεία είναι ο παράγοντας που ενσωματώνει την φυσική, ψυχολογική ή ακόμα και την πνευματική κατάσταση ενός ζώντος οργανισμού. Σύμφωνα με τον ορισμό που διατυπώθηκε στο καταστατικό του Παγκόσμιου Οργανισμού Υγείας η υγεία είναι η κατάσταση της πλήρους σωματικής, ψυχικής και κοινωνικής ευεξίας και όχι μόνο η απουσία ασθένειας ή αναπηρίας. Έτσι λοιπόν, η έννοια της υγείας, δεν αποδίδεται μόνο από την ιατρική, αλλά και από άλλους παράγοντες όπως είναι το περιβάλλον, η οικονομία, η εργασία κ.α.

Είναι επίσης σημαντικό να αναφερθεί ότι σε μία έρευνα όπου ζητήθηκε από συμμετέχοντες να κρίνουν τη σημαντικότητα διαφορετικών χαρακτηριστικών των ζωών τους, η καλή υγεία είχε τη μεγαλύτερη βαθμολογία (Campbell κ.α., 1976).

Η υγεία μπορεί να εκτιμηθεί αντικειμενικά και υποκειμενικά (Diener κ.α., 1989). Βέβαια η διάκριση μεταξύ αντικειμενικής κατάστασης της υγείας και υποκειμενικής αντίληψης για την υγεία δεν είναι πάντοτε ξεκάθαρη. Η αντικειμενική κατάσταση της υγείας είναι η κατάσταση της υγείας ενός ατόμου, όπως μπορεί να αναλυθεί από έναν ειδικό (γιατρό). Όμως η υποκειμενική αντίληψη κάθε ατόμου σχετικά με την κατάσταση της υγείας του, συνοψίζει τόσο την αντικειμενική κατάσταση όσο και την προσαρμογή του ατόμου σε αυτήν (Schimmel, 2009).

2.3 Ευημερία- Well-being

Η ευημερία (Well-being) είναι η κατάσταση στην οποία βρίσκεται ένα άτομο ή μια ομάδα, για παράδειγμα η κοινωνική, οικονομική, ψυχολογική, πνευματική κατάσταση ή κατάσταση υγείας (Diener κ.α., 1999). Ένα ψηλό επίπεδο ευημερίας σημαίνει ότι κατά κάποιο τρόπο η αίσθηση για την κατάσταση του ατόμου ή της ομάδας είναι θετική. Δηλαδή η ευημερία αναφέρεται στις ποικίλες και αλληλοσυνδεόμενες διαστάσεις της φυσικής, πνευματικής και κοινωνικής ευημερίας που εκτείνονται πέρα από τον παραδοσιακό ορισμό της υγείας. Αν και η ευημερία θεωρείται ότι συμπεριλαμβάνει και την υγεία, μερικοί άνθρωποι μπορεί να έχουν ικανοποιητικά επίπεδα ευημερίας παρόλο που έχουν ακόμη και συμπτώματα ψυχολογικής διαταραχής (Schimmel, 2009).

Η Ευημερία (WB) διαχωρίζεται στην Αντικειμενική (Objective WB), η οποία καθορίζεται βάσει αντικειμενικών κανόνων που ορίζουν την καλή ζωή, και στην Υποκειμενική (Subjective WB), η οποία ορίζεται ως "η συνολική εκτίμηση της ποιότητας της ζωής ενός ατόμου σύμφωνα με τα δικά του κριτήρια" (Shin και Johnson, 1978). Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (1997), η ποιότητα της ζωής είναι η αντίληψη ενός ατόμου σε σχέση με τους στόχους, τις προσδοκίες, τα πρότυπα και τις ανησυχίες του. Είναι μια ευρεία έννοια που επηρεάζεται από τη σωματική υγεία του ατόμου, την ψυχολογική του κατάσταση, τις προσωπικές του πεποιθήσεις και τις κοινωνικές του σχέσεις. Συχνά υποστηρίζεται ότι η Υποκειμενική Ευημερία αποτελείται από αλληλένδετες συνιστώσες (Dodge κ.α., 2012):

- την Ικανοποίηση από τη Ζωή (Life Satisfaction),
- το Θετικό και το Αρνητικό Συναίσθημα (Positive Affect & Negative Affect) και
- την Ευτυχία (Hapiness).

Η Ικανοποίηση από τη Ζωή βασίζεται σε αξιολογικές πεποιθήσεις (στάσεις) για τη ζωή ενός ατόμου. Είναι μια νοητική, κριτική διαδικασία και μπορεί να νοηθεί ως μια αξιολόγηση της ποιότητας ζωής ενός ανθρώπου σύμφωνα με τα δικά του κριτήρια. Δεν αποτελεί απλή αντανάκλαση των πραγματικών γεγονότων και συνθηκών αλλά

αντανακλά τη συναισθηματική διάθεση και τη γνωστική εκτίμηση που κάνει το άτομο για τα γεγονότα και τις συνθήκες ζωής. Σύμφωνα με τους Diener κ.α., (1985), η εκτίμηση του πόσο ικανοποιημένοι είναι οι άνθρωποι με την τωρινή κατάσταση τους βασίζεται σε ένα μέτρο σύγκρισης το οποίο θέτει ο κάθε άνθρωπος για τον εαυτό του χωρίς να καθορίζεται εξωτερικά. Το Θετικό και Αρνητικό Συναίσθημα από την άλλη, αφορά στην αξιολόγηση της Υποκειμενικής Ευημερίας με βάση το πόσο ευχάριστα και δυσάρεστα συναισθήματα βιώνουν οι άνθρωποι στη ζωή τους. Επομένως, μπορεί να ειπωθεί ότι η Ευημερία αποτελείται από δύο διακριτές συνιστώσες: ένα τμήμα, το οποίο αναφέρεται τόσο στην παρουσία θετικού συναισθήματος όσο και στην απουσία αρνητικού συναισθήματος, και ένα γνωστικό/νοητικό τμήμα (Diener κ.α., 1985). Η Ευτυχία τέλος, συχνά αναφέρεται ως Ευημερία και συγκεκριμένα ως ο συνδυασμός της Ικανοποίησης από τη Ζωή και της κατάστασης στην οποία το άτομο βιώνει περισσότερα θετικά παρά αρνητικά συναισθήματα. Παρά το γεγονός ότι οι έννοιες αυτές αναφέρονται σε διαφορετικές καταστάσεις, η στενή εννοιολογική σύνδεση ανάμεσα στην Ευημερία, στην Ικανοποίηση από τη Ζωή, στο Θετικό και Αρνητικό Συναίσθημα και στην Ευτυχία έχει οδηγήσει στην εννοιολογική τους ταύτιση. Στην παρούσα εργασία εξετάζονται μόνο οι εργασίες οι οποίες αντιμετωπίζουν τις παραπάνω έννοιες ως ταυτόσημες και για το λόγο αυτό η κάθε μία μπορεί να χρησιμοποιηθεί αντί της άλλης. Υπάρχουν μάλιστα έρευνες οι οποίες καλούνται "Happiness studies", και ασχολούνται με το εξειδεικευμένο ζήτημα της ευημερίας (Layard, 2005).

2.4 Σχέση υγείας και ευημερίας

Γενικά, θετικές καταστάσεις ευημερίας σχετίζονται με καλή κατάσταση υγείας (π.χ., Hilleras κ.α., 1998; Murrell κ.α., 2003; Ostir κ.α., 2000). Το πρόγραμμα ανάπτυξης των Ηνωμένων Εθνών (UNDP, 2017) αναφέρει χαρακτηριστικά ότι η καλύτερη υγεία αυξάνει την ατομική ευτυχία, κάτι που επιβεβαιώνεται στις "happiness studies" (Schimmel, 2009). Επιπλέον ο Wilson (1967) συμπέρανε ότι η υγεία είναι ισχυρά συσχετισμένη με την υποκειμενική ευημερία.

Ωστόσο οι παραπάνω διαπιστώσεις, θεωρείται ότι ισχύουν ως επί το πλείστον στις υποκειμενικές αντιλήψεις για την υγεία (George & Landerman, 1984; Larson, 1978; Okun, κ.α., 1984). Η συσχέτιση υγείας και ευημερίας εξασθενίζει όταν εξετάζονται αξιολογήσεις της υγείας από γιατρούς (Watten, κ.α., 1997).

Από τα παραπάνω, κρίθηκε αναγκαίο να αναφερθεί χωριστά η σχέση της ευημερίας με την αντικειμενική και υποκειμενική κατάσταση της υγείας, καθώς φαίνεται ότι συσχετίζονται με διαφορετικό τρόπο με την ευημερία.

- Η αντικειμενική κατάσταση υγείας είναι θετικά συσχετισμένη με την ατομική ευτυχία. Αρρώστιες που μπορεί να καταλήξουν και σε θάνατο, όπως ο καρκίνος, το HIV/AIDS, η ηπατίτιδα και ασθένειες που διαταράσσουν σημαντικά μια ομαλή λειτουργία του ανθρώπινου σώματος, γενικά έχουν αρνητική επίδραση στην ευτυχία (Diener και Seligman 2004, σ. 13). Παρόλα

αυτά αυτή η συσχέτιση είναι χαμηλή, πιθανόν λόγω της προσαρμογής του ανθρώπου σε αυτές τις ασθένειες (Layard 2005, σ. 69).

- Αντίθετα η υποκειμενική αντίληψη για την κατάσταση της υγείας, συνδέεται με τρόπο στατιστικά σημαντικό με το αίσθημα της ευτυχίας, ‘’καθώς συνοψίζει τόσο την αντικειμενική κατάσταση όσο και την προσαρμογή του ατόμου σε αυτήν ‘’ (Schimmel 2009, σ.101). Οι νευρωτικοί άνθρωποι, για παράδειγμα, αναφέρουν περισσότερα αναπνευστικά προβλήματα, αφού έχει διαγνωστεί η ασθένειά τους, από ότι στο παρελθόν (Diener κ.α. 1999, σ. 287). Αυτή η μεταβλητή περιλαμβάνει και τη φυσική κατάσταση και την συναισθηματική προσαρμογή, πράγμα το οποίο εξηγεί γιατί ένα άτομο αισθάνεται καλά, παρά το γεγονός ότι πάσχει από μια σοβαρή ασθένεια, και γιατί αισθάνεται άσχημα αν και αντικειμενικά έχει καλή υγεία.

Παράλληλα θα πρέπει να λάβουμε υπόψη μας κατά την ανάλυση, ότι η σχέση που συνδέει την υγεία και την ευτυχία είναι αμφίδρομη (Στρατουδάκη, 2011). Δεν επηρεάζει μόνο η υγεία την ευτυχία, αλλά και το αντίστροφο. Άτομα που αισθάνονται ευτυχισμένα και συνεπώς αισιόδοξα και με περισσότερη αυτοπεποίθηση, αντεπεξέρχονται καλύτερα στα προβλήματα υγείας και ζουν περισσότερο από άτομα που αισθάνονται λιγότερο ευτυχισμένα (Schimmel 2009, σ.102; Layard 2005, σ.23; Diener και Sheligman 2004, σ.14).

2.5 ΠΑΑ

Η παραγοντική ανάλυση των αντιστοιχιών, η οποία είναι η πλέον διαδεδομένη και πιο σημαντική μεθοδολογία από τις μεθόδους της ανάλυσης δεδομένων, της οικογένειας των πολυπαραγοντικών μεθόδων, έχει ως κύριο χαρακτηριστικό της, την απουσία a priori παραδοχής ότι τα δεδομένα ακολουθούν μια θεωρητική κατανομή (Benzecri, 1992). Με αυτόν τον τρόπο, τα δεδομένα ερμηνεύονται μόνο από τη συμπεριφορά που επιδεικνύουν και μια σημαντική έκφραση αυτής της ερμηνείας είναι η γραφική τους αναπαράσταση. Εφαρμόζεται σε δεδομένα κατηγορικών μεταβλητών, αλλά επίσης σε διατεταγμένα (όπως κλίμακες αξιολόγησης) και σε ποσοτικές μεταβλητές, οι οποίες μπορούν να μετασχηματιστούν σε κατηγορικές. Βασικές προϋποθέσεις εφαρμογής της μεθόδου, είναι η ύπαρξη ενός πίνακα θετικών αριθμών και επιπρόσθετα ότι τα προφίλ των γραμμών ή στηλών πρέπει να έχουν νόημα (Lebart κ.α., 2000). Πιο συγκεκριμένα, πρέπει οι τιμές των μεταβλητών να μην είναι αρνητικές και τα προφίλ (που είναι οι σχετικές συχνότητες) να μπορούν να υπολογιστούν. Για να επιτύχει η μέθοδος τη γραφική αναπαράσταση της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών, χρησιμοποιεί στοιχεία από τη γραμμική άλγεβρα και τη γεωμετρία, εφαρμόζοντας τεχνικές αλλαγής και ελάττωσης των διαστάσεων του αρχικού διανυσματικού χώρου των δεδομένων. Για τη μέτρηση της απόστασης ανάμεσα στα σημεία, χρησιμοποιείται η μετρική X^2. Τα ζητούμενα της ανάλυσης με την ΠΑΑ είναι: η εγγύτητα των σημείων γραμμών, η εγγύτητα των σημείων στηλών, η εγγύτητα των σημείων και οι σχέσεις σύνδεσης μεταξύ γραμμών και στηλών (Lebart κ.α., 1984).

Ορισμένα σημαντικά πλεονεκτήματα που προκύπτουν από την εφαρμογή της, είναι τα εξής:

- Το τελικό αποτέλεσμα παρουσιάζεται και με τη μορφή μιας εικόνας (γραφικής παράστασης), δίνοντας έτσι τη δυνατότητα να εξαχθούν, με πιο εύκολο τρόπο, χρήσιμα συμπεράσματα για τη συμπεριφορά των δεδομένων. Λόγω αυτής της ιδιότητάς της, θεωρείται από αρκετούς ερευνητές και ως μια περιγραφική μέθοδος ανάλυσης των δεδομένων.
- Η επιτακτική ανάγκη για έρευνα που κυριαρχεί στη σύγχρονη οικονομία, ιδιαίτερα στο πλαίσιο των κοινωνικών επιστημών όπου υπάρχει μια συνεχής ροή πληροφοριών, καθιστά την ΠΑΑ ένα χρήσιμο εργαλείο, καθώς μπορεί να αναλύσει ακόμη και πολύ μεγάλο όγκο δεδομένων.
- Τα δεδομένα που επεξεργάζονται με την ΠΑΑ έχουν κυρίως πολυδιάστατο χαρακτήρα και καταγράφονται σε πίνακες πολλών διαστάσεων. Αυτό το χαρακτηριστικό, της δίνει πολλά πλεονεκτήματα σε σχέση με τις μεθόδους της κλασικής στατιστικής.
- Όπως αναφέρεται στους Lebart κ.α. (2000), το αποτέλεσμα της εφαρμογής της ΠΑΑ έχει ποιοτικό όφελος, λόγω του ότι το στατιστικό υλικό μετά την ανάλυση έχει αναχθεί στα δομικά του χαρακτηριστικά, αλλά και ποσοτικό διότι έχει συνοψιστεί ολόκληρη η πληροφορία.

3.ΜΕΘΟΔΟΛΟΓΙΑ

Κύριος σκοπός της εργασίας ήταν η αναζήτηση των σημαντικότερων παραγόντων που επηρεάζουν την υποκειμενική αντίληψη για την υγεία. Με τη βοήθεια της Παραγοντικής ανάλυσης των αντιστοιχιών, επιδιώκεται η ανάλυση των δεδομένων της έρευνας ESS και η ανακάλυψη των μεταβλητών που φάνηκε ότι επηρεάζουν περισσότερο την υποκειμενική αντίληψη για την υγεία. Τα δεδομένα επεξεργάστηκαν και αναλύθηκαν με τη βοήθεια των λογισμικού SPSS και M.A.D.

Η ερώτηση για την υποκειμενική κατάσταση της υγείας ήταν η παρακάτω:

- ‘’Ποια είναι η κατάσταση της υγείας σας γενικά; ’’ Τιμές μεταβλητής: 1: πολύ καλή-5:πολύ κακή.

Τα κύρια χαρακτηριστικά των αποτελεσμάτων της ΠΑΑ και τα οποία χρησιμοποιούνται στις αναλύσεις που ακολουθούν, είναι τα παρακάτω:

- Η αδράνεια στην ΠΑΑ, είναι ένα μέτρο της απόστασης των προφίλ γραμμών (ή στηλών) από το μέσο προφίλ τους. Διαφορετικά, είναι η έκφραση του ποσοστού διασποράς των σημείων-προφίλ από το κέντρο βάρους τους, άρα όσο μεγαλύτερο είναι αυτό το ποσοστό, τόσο περισσότερο ξεχωρίζουν οι τιμές μεταξύ τους (δηλαδή τόσο μεγαλύτερη είναι και η διασπορά των σημείων στο χώρο).

- Δείκτης ερμηνείας (CTR): είναι ο δείκτης που διατυπώνει κατά πόσο το σημείο που αναλύεται, συμβάλλει στη δημιουργία του άξονα. Επιλέγονται ως σημαντικότερα, τα σημεία με μεγαλύτερο CTR.
- Το πρώτο παραγοντικό επίπεδο: Είναι δύο ευθείες κάθετες μεταξύ τους, που καταγράφουν ταυτόχρονα τη σύνθεση δύο τάσεων (γιατί συμπεριλαμβάνουν τον 1ο και 2ο παραγοντικό άξονα).
- Τα κύρια ζητούμενα της ανάλυσης, είναι η εγγύτητα των σημείων γραμμών και στηλών. Επομένως σημεία που βρίσκονται κοντά, σημαίνει ότι έχουν παρόμοια χαρακτηριστικά σε σχέση με τις μεταβλητές που αναπαριστούν.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Περιγραφικά στατιστικά

Ο Πίνακας 1 απεικονίζει το ποσοστό των ατόμων ανά κλάση της μεταβλητής ‘’Υποκειμενική κατάσταση της υγείας’’.

Πίνακας 1. *Κατανομή συχνοτήτων υποκειμενικής αντίληψης για την υγεία*

Υποκειμενική κατάσταση υγείας	**Απολ.συχνότητα**	**Σχετική συχνότητα**
Πολύ καλή	9.727	24,24
Καλή	17.059	42,50
Μέτρια	10.251	25,54
Κακή	2.534	6,31
Πολύ κακή	565	1,4
Σύνολο	**40.136**	**100**

Από τον παραπάνω πίνακα είναι σαφές, ότι περίπου 65% των ατόμων έχουν δηλώσει ότι έχουν καλή και πολύ καλή κατάσταση υγείας, κάτι που σημαίνει ότι η κατάσταση υγείας των ευρωπαίων πολιτών είναι σε αρκετά καλό επίπεδο.

4.2 Ανάλυση με την ΠΑΑ

Από την ανάλυση που προηγήθηκε με όλες τις μεταβλητές του ερωτηματολογίου, διαπιστώθηκε ότι οκτώ μεταβλητές επηρέασαν περισσότερο την υποκειμενική αντίληψη για την υγεία. Αρχικά παρουσιάζονται κάποια συγκεντρωτικά αποτελέσματα και μετέπειτα παρουσιάζεται ενδεικτικά η ανάλυση μίας μόνο μεταβλητής (της συχνότητας κοινωνικών επαφών).

Ακολουθεί ο Πίνακας 2 που παρουσιάζει την αδράνεια των αναλύσεων των οκτώ μεταβλητών (αυτών που διαπιστώθηκε ότι επηρεάζουν περισσότερο την και της υποκειμενικής αντίληψης για την υγεία.

Πίνακας 2. *Αδράνεια των αναλύσεων οκτώ μεταβλητών και της υποκειμενικής αντίληψης για την υγεία*

Μεταβλητές	**Αδράνεια**		
	Συνολική	**% 1ου άξονα**	**% 2ου άξονα**
Κοινωνικές επαφές	0,61	78,6	18,4
Ικανοποίηση από τη ζωή	0,16	79,8	16,6
Ευτυχισμένος	0,17	77,2	19,1
Ασφάλεια	0,08	80,2	18
Κατανάλωση φρούτων	0,01	72,5	24,4
Συχνότητα καπνίσματος	0,01	73,7	23,5
Συνολικά άτομα στο νοικοκυριό	0,05	90,8	7,9
Εισόδημα	0,07	95,5	3,5

Ακολουθεί η ανάλυση που έγινε με τη μέθοδο ΠΑΑ, για τη μεταβλητή **'**Πόσο συχνά συναντάτε σε κοινωνικές εκδηλώσεις φίλους, συγγενείς ή συναδέλφους σας;'

- Η ερώτηση αυτή έχει τιμές από 1:Ποτέ—7 Κάθε μέρα

Ο παρακάτω Πίνακας 3, παρουσιάζει τους δείκτες ερμηνείας της τρέχουσας ανάλυσης. Τα CTR που έχουν μεγαλύτερη τιμή από το μέσο όρο (που εδώ είναι 1000/5=250) και θεωρούνται πιο σημαντικά, παρουσιάζονται με έντονα γράμματα.

Πίνακας 3. *Δείκτες ερμηνείας της ανάλυσης των μεταβλητών της κοινωνικής επαφής και της υποκειμενικής αντίληψης για την υγεία*

Υποκ.κατ.υγείας	**1ος άξονας**		**2ος άξονας**	
Τιμή μεταβλ.	Συντεταγμ.	CTR	Συντεταγμ.	CTR
Πολύ καλή	0,440	0,208	-0,418	**0,387**
Καλή	0,170	0,056	0,122	0,06
Μέτρια	-0,309	0,110	0,329	**0,258**
Κακή	-1,075	**0,329**	-0,276	0,045
Πολύ κακή	-2,166	**0,297**	-1,384	0,251

Από την επιλογή των σημαντικότερων σημείων με βάση το δείκτη CTR, ακολουθούν τα σημαντικότερα σημεία ανά άξονα στους Πίνακες 4 και 5.

Πίνακας 4. *Σημαντικότερα σημεία ανά άξονα -1ος άξονας*

Μεταβλητές	**Συντεταγμένες**	**CTR**
Κακή(Υποκ.κατ.υγ.4)	-1,075	0,329
Πολύ κακή (Υποκ.κατ.υγ.5)	-2,166	0,297
Ποτέ (Πόσο συχνά συναντ1)	-2,351	0,480
Λιγότερο από μήνα(Πόσο συχνά συναντ2)	-0,945	0,328

Στον ανωτέρω πίνακα που περιέχει τα σημαντικότερα σημεία του 1ου άξονα, είναι σαφές ότι τα άτομα με κακή υποκειμενική κατάσταση υγείας, έχουν λιγότερες κοινωνικές επαφές, καθώς τα σημεία που απεικονίζουν αυτές τις τιμές των μεταβλητών είναι κοντά, όπως φαίνεται και από τις συντεταγμένες τους.

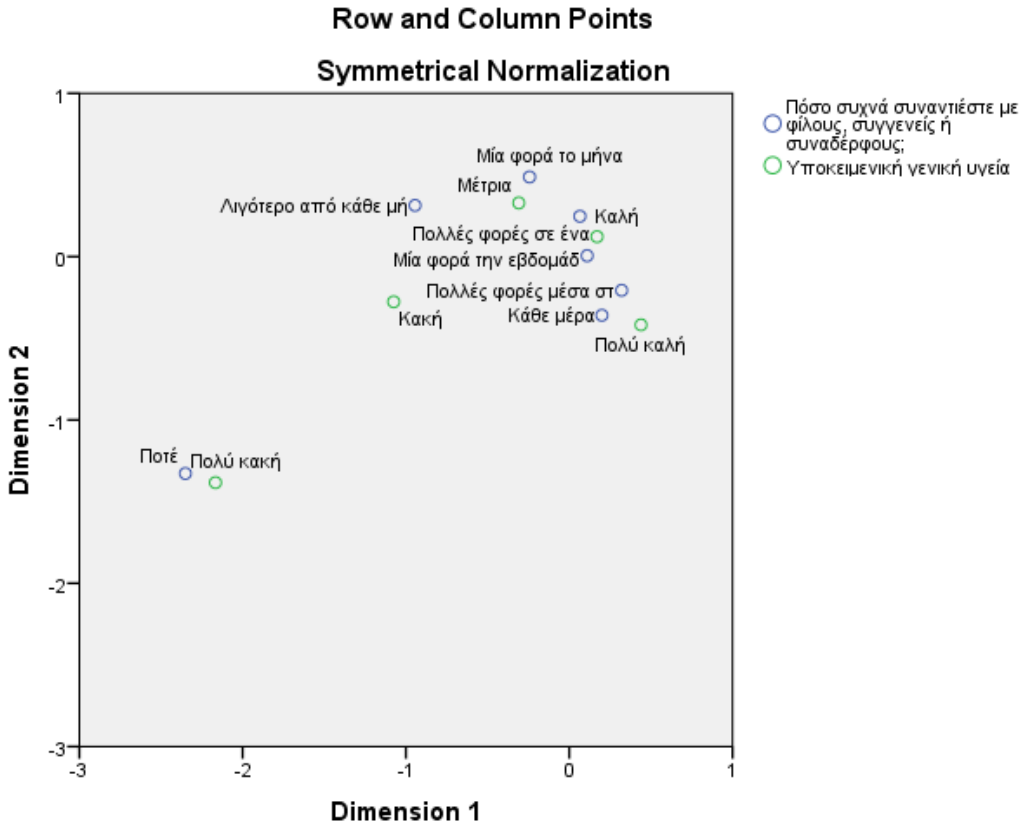
Πίνακας 5. *Σημαντικότερα σημεία ανά άξονα-2ος άξονας*

Μεταβλητές	**Συντεταγμένες**	**CTR**
Πολύ κακή(Υποκ.κατ.υγ.5)	-0,418	0,387
Μέτρια(Υποκ.κατ.υγ.3)	0,329	0,258
Ποτέ (Πόσο συχνά συναντ1)	-1,328	0,316
Μία φορά το μήνα(Πόσο συχνά συναντ3)	0,488	0,227

Στον Πίνακα 5 που απεικονίζει τα σημαντικότερα σημεία του 2ου άξονα, παρατηρείται ότι τα άτομα που δήλωσαν ότι είχαν πολύ κακή κατάσταση υγείας, δε συναντήθηκαν με κανέναν, ενώ όσοι είχαν λίγες σχετικά συναντήσεις δήλωσαν ότι είχαν μέτρια υγεία. Τα αποτελέσματα αυτά είναι ανάλογα με τον 1ο άξονα.

Το 1ο παραγοντικό διάγραμμα ακολουθεί στην Εικόνα 1. Στο διάγραμμα αυτό φαίνεται ότι υπάρχει μια μεγάλη **αλληλεξάρτηση** των μεταβλητών, η οποία προκύπτει από την εγγύτητα συγκεκριμένων σημείων. Πιο συγκεκριμένα, αυτοί που δηλώνουν ότι συναντιούνται συχνότερα με γνωστούς ή φίλους, δηλώνουν ότι έχουν και καλύτερη υγεία. Από την άλλη, αυτοί που δε συναντιούνται ποτέ με άλλους, δηλώνουν ότι έχουν πολύ κακή υγεία.

Εικόνα 1. *1ο παραγοντικό επίπεδο*



Οι σημαντικότερες μεταβλητές-παράγοντες που επηρέασαν την υποκειμενική υγεία, παρουσιάζονται στον Πίνακα 6 με τη μορφή ερωτήσεων και απαντήσεων.

Πίνακας 6. *Σημαντικότερες μεταβλητές για τον επηρεασμό της υποκειμενικής κατάστασης της υγείας*

Ερώτημα	**Ποιοι έχουν καλύτερη υποκ.υγεία**
Κάνοντας μια γενική εκτίμηση, πόσο ικανοποιημένος/η είστε γενικά με τη ζωή σας σήμερα;	Αυτοί που είναι πιο ικανοποιημένοι από τη ζωή τους
Πόσο συχνά συναντάτε σε κοινωνικές εκδηλώσεις φίλους, συγγενείς ή συναδέλφους σας;	Οσοι συναντιούνται πιο συχνά με άλλους

Πόσο ασφαλής νιώθετε – ή θα νιώθατε – περπατώντας μόνος/η σας στη γειτονιά σας όταν σκοτεινιάσει;	Οσοι αισθάνονται πιο ασφαλείς στη γειτονιά τους
Πόσο συχνά καταναλώνεται φρούτα και λαχανικά;	Οσοι καταναλώνουν περισσότερα φρούτα και λαχανικά
Πόσα τσιγάρα καταναλώνετε;	Οσοι καπνίζουν λιγότερα τσιγάρα
Πόσα άτομα μένουν συνολικά στο ίδιο νοικοκυριό;	Οσοι μένουν με περισσότερα άτομα
Ποιο είναι το συνολικό εισόδημα όλων των μελών του νοικοκυριού;	Οσοι έχουν μεγαλύτερο εισόδημα
Σε γενικές γραμμές, πόσο ευτυχισμένος/η θα λέγατε ότι είστε;	Οσοι νιώθουν πιο ευτυχισμένοι

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα έρευνα χρησιμοποίησε δεδομένα από μια ευρωπαϊκή έρευνα, που πραγματοποιήθηκε σε ένα δείγμα 40185 κατοίκων ευρωπαϊκών χωρών. Οι μεταβλητές-ερωτήσεις της έρευνας ήταν κυρίως οικονομικού, κοινωνικού και πολιτικού περιεχομένου. Βασικός σκοπός ήταν να ερευνηθούν οι σημαντικότεροι παράγοντες που επηρεάζουν την υποκειμενική κατάσταση της υγείας των πολιτών αυτών. Από την παραπάνω ανάλυση, διαπιστώθηκε ότι ορισμένοι παράγοντες σχετίζονται άμεσα με την υποκειμενική αντίληψη για την υγεία των ευρωπαίων πολιτών.

Ήταν διακριτό ότι διατροφικές συνήθειες όπως η μεγαλύτερη κατανάλωση φρούτων και το λιγότερα συχνό κάπνισμα, ευνοούν τη διαμόρφωση καλύτερης υγείας. Η πιο υγιεινή διατροφή είναι ένας παράγοντας που βέβαια επηρεάζει την αντικειμενική κατάσταση της υγείας, αλλά διαπιστώθηκε ότι και η υποκειμενική κατάσταση επηρεάστηκε σε μεγάλο βαθμό.

Επιπλέον άνθρωποι που δήλωσαν ότι είναι πιο ικανοποιημένοι από τη ζωή τους και πιο ευτυχισμένοι, δήλωσαν ότι είχαν και καλύτερη υγεία. Αυτό είναι λογικό και υπάρχει μια αμφίδρομη σχέση μεταξύ υγείας και ευτυχίας όπως προαναφέρθηκε στην παράγραφο 2.4.

Η κοινωνική ένταξη των ατόμων είναι ένας ακόμη παράγοντας που συντελεί στη διαμόρφωση καλύτερης κατάστασης υγείας, όπως προέκυψε από την ανάλυση των

δεδομένων αυτής της έρευνας. Έχει αναφερθεί από τον Elo (2009), ότι τα άτομα που έχουν ευρύτερο κοινωνικό δίκτυο, ευνοούνται σε ψυχολογικό επίπεδο και αυτό είναι κάτι που βελτιώνει και την υποκειμενική αντίληψη για την υγεία. Σε πολλές έρευνες υγείας έχει αναφερθεί ότι αρκετοί παράγοντες που σχετίζονται με την αντίληψη για της υγείας είναι ψυχολογικοί (Gataulinas & Bancevica, 2014).

Η ασφάλεια που νιώθουν οι άνθρωποι στην περιοχή-γειτονιά που μένουν, ήταν άλλος ένας παράγοντας που επηρέασε σημαντικά την υποκειμενική αντίληψη για την υγεία τους. Πιο συγκεκριμένα τα άτομα που αισθάνονταν πιο ασφαλή, είχαν καλύτερη αντίληψη για την κατάσταση της υγείας τους.

Το ψηλότερο εισόδημα ευνοεί τη βελτίωση της υποκειμενικής υγείας και αυτό έχει καταγραφεί σε αρκετές έρευνες (Prag κ.α., 2013). Είναι σαφές ότι άτομα με μεγαλύτερο εισόδημα, μπορούν να εξασφαλίσουν καλύτερες συνθήκες περίθαλψης και αντιμετώπισης των ασθενειών. Σε αυτήν την εργασία το αποτέλεσμα ήταν παρεμφερές, αφού τα άτομα με μεγαλύτερο συνολικό οικογενειακό εισόδημα, δήλωσαν ότι έχουν καλύτερο επίπεδο υγείας, σε σχέση με εκείνους που είχαν χαμηλότερα εισοδήματα.

ABSTRACT

Health is considered a significant factor that can affect seriously citizens' well-being, which the governments of all countries deal with great seriousness and concern. An important subject is which factors affect the state of health of the citizens. In this study there are examined mostly social factors, which can be proved crucial for the definition of the state of subjective perception for health. For this purpose, data from a big social survey that is conducted every two years in Europe, is analysed. This survey is European Social Survey (ESS), which provides many important aspects for the social, economic and cultural life of European citizens. From this study, the most important social, economic and cultural factors were researched, in order to discover which affect the subjective perception for health. To achieve this, a statistical methodology was utilized, Correspondence analysis, which has the potential to designate easily the differentiations between the variables or/and their values. The results proved that the most determinant factors affecting health are: income, depression, general satisfaction and happiness. It was also derived that some other factors do not influence health, as marital status, religion, domicile etc.

Keywords: European social survey, Well-being, Health, Subjective perception for health, Correspondence analysis, Social-economic-cultural factors.

ΑΝΑΦΟΡΕΣ

Benzecri J.P., (1992). *Correspondence Analysis Handbook,* New York: Taylor & Francis.

Campbell, A., Converse, P.E., & Rodjers, W.L. (1976). The quality of American life, New York: Russell Sage Foundation.

Diener, E., et al. (1999). Subjective well-being: Three decades of progress. Psychological Bulletin, Vol. 125, Iss: 2, pp.276–302.

Diener, E., & Seligman, M. E. P. (2004). 'Beyond money—toward an economy of well-being'. Psychological Science in the Public Interest, Vol.5, Iss:1, pp.1–31.

Diener, E. Sandvik, E., & Larsen, R.J. (1985). Age and sex effects for emotional intensity. Developmental Pscyhology, Vol.21, pp.542-546.

Dodge, R., Daly, A., Huyton, J., & Sanders, L. (2012). The challenge of defining wellbeing. International Journal of Wellbeing, 2(3), pp.222-235.

Elo, I. T. (2009). Social class differentials in health and mortality. Patterns and explanations in comparative perspective. Annual Review of Sociology, Vol.35, pp.553–572.

ESS Round 7: European Social Survey Round 7 Data (2014). Data file edition 2.1. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

Fotso, J. C., & Kuate-Defo, B. (2005). Measuring Socioeconomic Status in Health Research in Developing Countries: Should We Be Focusing on Households, Communities or Both? Social Indicators Research, 72, 189-237. http://dx.doi.org/10.1007/s11205-004-5579-8

Gataulinas A. & Mancevica M., (2014), Subjective health and subjective well-being (the case of EU countries), Advances in Applied Sociology, Vol.4, pp.212-223.

George, L.K., & Landerman, R. (1984). Health and subjective well-being: a replicated secondary data analysis. International Journal of Aging and Human development, Vol.19, pp.133-156.

Hilleras, P.K., Jorm, A.F., Herlitz, A., & Winblad, B. (1998). Negative and positive affect among the very old: A survey on a sample age 90 years or older. Research on Aging, Vol. 20, pp.593–610.

Layard, R. (2005). Happiness—lessons from a New Science. New York: Penguin Press.

Lebart L., Morineau A., Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis*, New York, Wiley.

Lebart L., Morineau A., Piron M. (2000). *Statistique Exploratoire Multidimensionelle*, Paris, Dunod.

Murrell, S.A., Salsman, N.L., & Meeks, S. (2003). Educational attainment, positive psychological mediators, and resources for health and vitality in older adults. Journal of Aging and Health, Vol.15, pp.591–615.

Okun, M.A., Stock, W.A., Haring M.J., & Witter, R.A. (1984). Health and subjective well-being: a meta-analysis. International Journal of Aging and Human development, Vol.19, pp.111-132.

Ostir, G.V., Markides, K.S., Black, S.A., & Goodwin, J.S. (2000). Emotional well-being predicts subsequent functional independence and survival. Journal of the American Geriatrics Society, Vol.48, pp.473–478.

Prag P., Mills M., & Wittek R., (2013), Income and income inequality as social determinants of health: do social copmarisons paly a role ?, European Sociological Review, Vol.30, Num.2, pp.:218-229.

Schimmel, J. (2009), Development as Happiness: The Subjective Perception of Happiness and UNDP's Analysis of Poverty, Wealth and Development, Journal of Happiness Studies , DOI 10.1007/s10902-007-9063-4.

Shin D., and Johnson D., (1978). Avowed happiness as an overall assesment of the quality of life, Social Indicators Research, Vol.5, pp.475-492.

United nations development program, (2017), retrieved on-line 02/04/2017, http://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-3-good-health-and-well-being.html.

Watten, R.G., Vassend., D., Myhrer, T. & Syversen, J.L. (1997). Personality factors and somatic symptoms. European Journal of Personality, Vol.11, pp.57-68.

Wilson W. (1967). Correletaes of avowed happiness, Psychological Bulletin, Vol.6, pp.294-306.

World Health Organization. (1997). Measuring quality of life. Geneva: WHO (WHO/MSA/MNH/PSF/97.4).

Στρατουδάκη Χ. (2011). ''Υποκειμενική ευημερία: ευτυχία και ικανοποίηση από τη ζωή στην Ελλάδα'', στο Παπλιάκου Β., Σταθοπούλου Θ., Στρατουδάκη Χ. (επιμ.), Θεσμοί-Αξίες-Συμπεριφορές: Μελέτη των ευρημάτων της Ευρωπαϊκής Κοινωνικής Έρευνας, Αθήνα, ΕΚΚΕ.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 30ου Πανελληνίου Συνεδρίου Στατιστικής (2017), σελ. 35−43

ΚΑΤΑΝΟΜΗ ΤΟΥ ΕΛΑΧΙΣΤΟΥ ΚΑΙ ΤΟΥ ΜΕΓΙΣΤΟΥ ΑΠΟ ΤΥΧΑΙΟ ΑΡΙΘΜΟ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Παναγιώτης Μπομποτάς, Μάρκος Β. Κούτρας
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
{pbobotas, mkoutras}@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή γίνεται μελέτη της κατανομής του ελαχίστου και του μεγίστου όταν ο αριθμός των συνεχών και θετικών τυχαίων μεταβλητών του δείγματος είναι τυχαία μεταβλητή. Στην περίπτωση που το στήριγμα της κατανομής του αριθμού των τυχαίων μεταβλητών του δείγματος επιτρέπεται να περιέχει το μηδέν οδηγούμαστε σε προβλήματα τα οποία δεν έχουν μελετηθεί συστηματικά στη βιβλιογραφία. Στην παρούσα εργασία, δίνονται οι ροπές, οι συναρτήσεις επιβίωσης καθώς και οι συναρτήσεις κινδύνου των κατανομών που παράγονται. Τέλος, γίνεται εφαρμογή των αποτελεσμάτων σε συγκεκριμένες κατανομές.

Λέξεις κλειδιά: μοντέλα ανάλυσης επιβίωσης, κατανομές μικτού τύπου, Lehmann-type alternatives κατανομές, εμφυτευμένη σε Μαρκοβιανή αλυσίδα τυχαία μεταβλητή διωνυμικού τύπου

1. ΕΙΣΑΓΩΓΗ

Έστω $X_1, X_2, \ldots, X_N$ τυχαίο δείγμα ενός τυχαίου αριθμού συνεχών θετικών τυχαίων μεταβλητών από μία κατανομή F με πυκνότητα πιθανότητας f, δηλαδή το N είναι τυχαία μεταβλητή (θετικών ακεραίων) από διακριτή κατανομή με γεννήτρια πιθανότητας $P_N(z) = \mathsf{E}(z^N)$. Θέτουμε

$$X_{(1)} = \min\{X_1, X_2, \ldots, X_N\} \quad \text{και} \quad X_{(N)} = \max\{X_1, X_2, \ldots, X_N\}.$$

Στη βιβλιογραφία έχουν εμφανιστεί αρκετές εργασίες όπου έχουν μελετηθεί οι συναρτήσεις κατανομής του $X_{(1)}$ ή/και του $X_{(N)}$, για κατανομές του N με $\mathsf{P}(N = 0) = 0$. Σε αυτό το πλαίσιο ισχύει ότι

$$F_{X_{(1)}}(t) = 1 - P_N(1 - F(t)) \quad \text{και} \quad F_{X_{(N)}}(t) = P_N(F(t)),$$

αντίστοιχα. Ενδεικτικά αναφέρονται οι εργασίες των Adamidis and Loukas (1998), Kus (2007), Tahmasbi and Rezaei (2008), Chahkandi and Ganjali (2009), Barreto-Souza et al. (2011), Morais and Barreto-Souza (2011), Louzada et al. (2011), Louzada et al. (2012), Tojeiro et al. (2014), στις οποίες χρησιμοποιείται για την F είτε

η εκθετική, είτε η Weibull κατανομή, ενώ για την N χρησιμοποιούνται η γεωμετρική, η κόλουρη Poisson, η λογαριθμική, ή η κόλουρη power series κατανομή. Τα παραγόμενα μοντέλα στις παραπάνω εργασίες περιγράφουν προβλήματα, για παράδειγμα σε μοντέλα ανάλυσης επιβίωσης, όπου η αστοχία μιας συσκευής προκύπτει από την παρουσία ενός αγνώστου αριθμού ελαττωματικών εξαρτημάτων του ιδίου είδους. Η βασική υπόθεση που γίνεται στα μοντέλα αυτά είναι ότι για να υπάρξει αστοχία της συσκευής θα πρέπει να εμφανισθεί τουλάχιστον ένα ελαττωματικό εξάρτημα.

Από την άλλη πλευρά, σε προβλήματα προληπτικού ελέγχου συντήρησης συστημάτων, είναι δυνατόν να μην υπάρξει κάποιο εξάρτημα που θα χρειαστεί συντήρηση, οπότε και ο χρόνος συντήρησης του συστήματος είναι μηδέν. Το γεγονός αυτό μπορεί να ενσωματωθεί στο μοντέλο επιτρέποντας το στήριγμα της κατανομής της N να περιέχει το 0. Εάν υπάρξουν εξαρτήματα που χρειάζονται συντήρηση, $N > 0$, το $X_{(1)}$ είναι ο ελάχιστος χρόνος συντήρησης των εξαρτημάτων και το $X_{(N)}$ είναι ο μέγιστος χρόνος συντήρησης των εξαρτημάτων (άρα και ο μέγιστος χρόνος για την πλήρη επαναλειτουργία του συστήματος). Στην παρούσα εργασία γίνεται μελέτη των κατανομών του ελαχίστου και του μεγίστου όταν $\mathsf{P}(N = 0) \neq 0$. Προς αυτήν την κατεύθυνση ορίζουμε

$$T_1 = \begin{cases} 0, & N = 0, \\ X_{(1)}, & N > 0, \end{cases} \quad \text{και} \quad T_2 = \begin{cases} 0, & N = 0, \\ X_{(N)}, & N > 0, \end{cases} \tag{1}$$

για τις οποίες προκύπτουν οι μικτού τύπου κατανομές (mixed-type distributions)

$$F_{T_1}(t) = \begin{cases} \mathsf{P}(N = 0) > 0, & t = 0, \\ 1 - P_N(1 - F(t)) + \mathsf{P}(N = 0), & t > 0, \end{cases} \tag{2}$$

και

$$F_{T_2}(t) = \begin{cases} \mathsf{P}(N = 0) > 0, & t = 0, \\ P_N(F(t)), & t > 0, \end{cases} \tag{3}$$

με πυκνότητες πιθανότητας

$$f_{T_1}(t) = \begin{cases} \mathsf{P}(N = 0) > 0, & t = 0, \\ P'_N(1 - F(t))\, f(t), & t > 0, \end{cases} \tag{4}$$

και

$$f_{T_2}(t) = \begin{cases} \mathsf{P}(N = 0) > 0, & t = 0, \\ P'_N(F(t))\, f(t), & t > 0, \end{cases} \tag{5}$$

αντίστοιχα, με $P'_N(z) = dP_N(z)/dz$.

Στην Ενότητα 2 δίνονται τα κύρια αποτέλεσματα της εργασίας για την κατανομή του ελαχίστου και του μεγίστου, ενώ στην Ενότητα 3 γίνεται εφαρμογή των αποτελεσμάτων σε συγκεκριμένες κατανομές.

2. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Από τον ορισμό της γεννήτριας πιθανότητας της N ισχύει ότι $P_N(z) = \sum_{r=0}^{\infty} \alpha_r z^r$ (αν το στήριγμα της N είναι πεπερασμένο θα είναι πεπερασμένο και το πλήθος των όρων του αθροίσματος) με

$$\alpha_r = \mathsf{P}(N = r), \quad r = 0, 1, 2, \ldots, \quad \text{και } \sum_{r=0}^{\infty} \alpha_r = 1,$$

οπότε από τις σχέσεις (2) και (3), για $t > 0$, προκύπτει ότι

$$\begin{aligned} F_{T_1}(t) &= 1 - P_N(1 - F(t)) + \mathsf{P}(N = 0) = \sum_{r=0}^{\infty} \alpha_r - \sum_{r=0}^{\infty} \alpha_r (1 - F(t))^r + \alpha_0 \\ &= \alpha_0 + \sum_{r=1}^{\infty} \alpha_r G_r^{(1)}(t) \end{aligned} \tag{6}$$

και

$$\begin{aligned} F_{T_2}(t) &= P_N(F(t)) = \sum_{r=0}^{\infty} \alpha_r (F(t))^r \\ &= \alpha_0 + \sum_{r=1}^{\infty} \alpha_r G_r^{(2)}(t) \end{aligned} \tag{7}$$

όπου $G_r^{(1)}(t) = 1 - (1 - F(t))^r$ είναι η συνάρτηση κατανομής της ελάχιστης παρατήρησης τυχαίου δείγματος μεγέθους r από την κατανομή F και $G_r^{(2)}(t) = (F(t))^r$ είναι η συνάρτηση κατανομής της μέγιστης παρατήρησης τυχαίου δείγματος μεγέθους r από την κατανομή F. Στη βιβλιογραφία οι $G_r^{(1)}$ και $G_r^{(2)}$ αναφέρονται και ως Lehmann-type alternatives κατανομές, βλέπε Lehmann (1953), ενώ για τη συνέχεια, πιο συγκεκριμένα, θα αναφέρονται ως Lehmann alternatives κατανομές τύπου I και τύπου II , αντίστοιχα. Σημειώνεται ότι εάν για μία συνάρτηση κατανομής F θεωρήσουμε τον έλεγχο υποθέσεων με $\mathcal{H}_0 : X \sim F(x)$, οι οικογένειες κατανομών Lehmann alternatives τύπου I και τύπου II προκύπτουν ως οι κατανομές των εναλλακτικών υποθέσεων $\mathcal{H}_1 : X \sim 1 - (1 - F(x))^\rho$, με $\rho \neq 1$ και $\mathcal{H}_1 : X \sim (F(x))^\rho$, αντίστοιχα. Τέτοιες υποθέσεις χρησιμοποιούνται συνήθως στην περιοχή της Στατιστικής μοντελοποίησης και του Στατιστικού Ελέγχου Ποιότητος για τη μελέτη της ισχύος διάφορων μη-παραμετρικών διαγραμμάτων ελέγχου, βλέπε Gupta et al. (1998), Gibbons and Chakraborti (2003) και Hollander and Wolfe (1999).

Με βάση τις σχέσεις (4) και (5) και τη συζήτηση που ακολουθεί τις σχέσεις (6) και (7) σχετικά με τις Lehmann-type alternatives κατανομές προκύπτει το επόμενο αποτέλεσμα.

Πρόταση 1. *Εάν $T = T_1$ ή $T = T_2$ όπως στην (1), τότε η κατανομή του T είναι μία μικτού τύπου κατανομή με θετική συνάρτηση πιθανότητας στο 0, $f_T(0) = \mathsf{P}(N = 0)$, ενώ για $t > 0$, η $f_T(t)$ εκφράζεται ως γραμμικός συνδυασμός Lehmann-type alternatives κατανομών με θετικούς συντελεστές που αθροίζουν σε τιμή μικρότερη της μονάδας, δηλαδή*

$$f_T(t) = \sum_{r=1}^{\infty} \alpha_r g_r(t), \quad t > 0,$$

όπου

$$\alpha_r = \mathsf{P}(N = r), \quad r \geq 1, \quad \mu\epsilon \sum_{r=1}^{\infty} \alpha_r < 1,$$

$$g_r(t) = \begin{cases} r(1 - F(t))^{r-1} f(t), & \alpha\nu\ T = T_1, \\ r(F(t))^{r-1} f(t), & \alpha\nu\ T = T_2. \end{cases}$$

Στο σημείο αυτό, αξίζει να αναφερθεί ότι στην περίπτωση που η F είναι εκθετική κατανομή τότε και η Lehmann alternative τύπου I είναι επίσης εκθετική (με διαφορετική παράμετρο), ενώ στην περίπτωση που η F είναι κατανομή δύναμης τότε και η Lehmann alternative τύπου II είναι επίσης κατανομή δύναμης. Λαμβάνοντας υπόψη την τελευταία παρατήρηση προκύπτουν τα ακόλουθα αποτελέσματα.

Πόρισμα 1. *Έστω $X_1, X_2, \ldots$, μία ακολουθία ανεξάρτητων και ισοκατανεμημένων τυχαίων μεταβλητών από την εκθετική κατανομή $E(\lambda)$ με $F(t) = 1 - \exp\{-\lambda t\}$, $\lambda > 0$, $t > 0$. Τότε για την $T = T_1$ η $f_T(t)$, $t > 0$, εκφράζεται ως γραμμικός συνδυασμός πυκνοτήτων πιθανοτήτων εκθετικών κατανομών $E(r\lambda)$, $r \geq 1$, με θετικούς συντελεστές που αθροίζουν σε τιμή μικρότερη της μονάδας, δηλαδή*

$$f_T(t) = \sum_{r=1}^{\infty} \alpha_r (r\lambda \exp\{-r\lambda t\}), \quad \alpha_r = \mathsf{P}(N = r), \quad t > 0.$$

Πόρισμα 2. *Έστω $X_1, X_2, \ldots$, μία ακολουθία ανεξάρτητων και ισοκατανεμημένων τυχαίων μεταβλητών από την κατανομή δύναμης $Power(\lambda)$ με $F(t) = t^{\lambda}$, $\lambda > 0$, $0 < t < 1$. Τότε για την $T = T_2$ η $f_T(t)$, $0 < t < 1$, εκφράζεται ως γραμμικός συνδυασμός κατανομών δύναμης $Power(r\lambda)$, $r \geq 1$, με θετικούς συντελεστές που αθροίζουν σε τιμή μικρότερη της μονάδας, δηλαδή*

$$f_T(t) = \sum_{r=1}^{\infty} \alpha_r (r\lambda\, t^{r\lambda - 1}), \quad \alpha_r = \mathsf{P}(N = r), \quad 0 < t < 1.$$

Στη συνέχεια δίνονται αποτελέσματα σχετικά με τις ροπές και τις ροπογεννήτριες των T_1 και T_2, καθώς και τις συναρτήσεις επιβίωσης και τους ρυθμούς κινδύνου αυτών.

Πρόταση 2. *Εάν $T = T_1$ ή $T = T_2$ τότε :*

(α) $\mathsf{E}(T^k) = \sum_{r=1}^{\infty} \alpha_r \mu'_{k,r}$, *όπου* $\mu'_{k,r} = \mathsf{E}(Y_r^k)$, $Y_r \sim g_r(t)$, *δηλαδή η ροπή τάξης* k *της* T *εκφράζεται ως γραμμικός συνδυασμός ροπών τάξης* k *Lehmann-type alternatives κατανομών με θετικούς συντελεστές που αθροίζουν σε τιμή μικρότερη της μονάδας.*

(β) $\mathsf{E}(e^{zT}) = \sum_{r=0}^{\infty} \alpha_r M_{Y_r}(z)$, *όπου* $M_{Y_r}(z) = \mathsf{E}(e^{zY_r})$, $Y_r \sim g_r(t)$, $r \geq 1$ *και* $M_{Y_0}(z) = 1$.

Απόδειξη. (α) Λαμβάνοντας υπόψη την Πρόταση 1 ισχύει ότι

$$\mathsf{E}(T^k) = 0^k \mathsf{P}(T = 0) + \int_0^{\infty} t^k f_T(t) dt = \sum_{r=1}^{\infty} \alpha_r \int_0^{\infty} t^k g_r(t) dt = \sum_{r=1}^{\infty} \alpha_r \mu'_{k,r}.$$

(β) Εργαζόμενοι ομοίως όπως στο (α) προκύπτει $\mathsf{E}(e^{zT}) = \mathsf{P}(N = 0) + \sum_{r=1}^{\infty} \alpha_r M_{Y_r}(z)$ και αυτό ολοκληρώνει την απόδειξη. □

Πρόταση 3. *Εάν $T = T_1$ ή $T = T_2$ τότε :*

(α) Η συνάρτηση επιβίωσης της T δίνεται από τη σχέση

$$S_T(t) = \begin{cases} \mathsf{P}(N > 0) = \sum_{r=1}^{\infty} \alpha_r, & t = 0, \\ \sum_{r=1}^{\infty} \alpha_r (1 - F(t))^r, & \text{αν } T = T_1, \quad t > 0, \\ \sum_{r=1}^{\infty} \alpha_r (1 - (F(t))^r), & \text{αν } T = T_2, \quad t > 0, \end{cases}$$

(β) Η συνάρτηση κινδύνου της T δίνεται από τη σχέση

$$h_T(t) = \begin{cases} \dfrac{\mathsf{P}(N = 0)}{\mathsf{P}(N > 0)}, & t = 0, \\ f(t) \dfrac{\sum_{r=1}^{\infty} \alpha_r r (1 - F(t))^{r-1}}{\sum_{r=1}^{\infty} \alpha_r (1 - F(t))^r}, & \text{αν } T = T_1, \quad t > 0, \\ f(t) \dfrac{\sum_{r=1}^{\infty} \alpha_r r (F(t))^{r-1}}{\sum_{r=1}^{\infty} \alpha_r (1 - (F(t))^r)}, & \text{αν } T = T_2, \quad t > 0, \end{cases}$$

Απόδειξη. (α) Επειδή $S_T(t) = 1 - F_T(t)$, η απόδειξη είναι άμεση λαμβάνοντας υπόψη τις σχέσεις (2), (3), (6) και (7).
(β) Επειδή $h_T(t) = f_T(t)/S_T(t)$, η απόδειξη είναι άμεση λαμβάνοντας υπόψη την Πρόταση 1, και το (α). □

Για παράδειγμα, εάν $T = T_1$ και $F(t) = 1 - \exp\{-\lambda t\}$, $\lambda > 0$, $t > 0$, τότε από

τις Προτάσεις 2, 3 και το Πόρισμα 1 προκύπτει ότι

$$\begin{aligned}
&\mathsf{E}(T) = \lambda^{-1}\sum_{r=1}^{\infty}\alpha_r/r, \quad \mathsf{E}(T^2) = 2\lambda^{-2}\sum_{r=1}^{\infty}\alpha_r/r^2,\\
&\mathsf{Var}(T) = 2\lambda^{-2}\sum_{r=1}^{\infty}\alpha_r/r^2 - \left(\lambda^{-1}\sum_{r=1}^{\infty}\alpha_r/r\right)^2,\\
&h_T(t) = \begin{cases} \dfrac{\mathsf{P}(N=0)}{\mathsf{P}(N>0)}, & t=0,\\ \dfrac{\sum_{r=1}^{\infty}\alpha_r(r\lambda)\exp\{-r\lambda t\}}{\sum_{r=1}^{\infty}\alpha_r\exp\{-r\lambda t\}}, & t>0, \end{cases}
\end{aligned} \tag{8}$$

ενώ εάν $T = T_2$ και $F(t) = t^{\lambda}$, $\lambda > 0$, $0 < t < 1$, τότε από τις Προτάσεις 2, 3 και το Πόρισμα 2 προκύπτει ότι

$$\begin{aligned}
&\mathsf{E}(T) = \sum_{r=1}^{\infty}\alpha_r\frac{r\lambda}{r\lambda+1}, \quad \mathsf{E}(T^2) = \sum_{r=1}^{\infty}\alpha_r\frac{r\lambda}{r\lambda+2},\\
&\mathsf{Var}(T) = \sum_{r=1}^{\infty}\alpha_r\frac{r\lambda}{r\lambda+2} - \left(\sum_{r=1}^{\infty}\alpha_r\frac{r\lambda}{r\lambda+1}\right)^2,\\
&h_T(t) = \begin{cases} \dfrac{\mathsf{P}(N=0)}{\mathsf{P}(N>0)}, & t=0,\\ \dfrac{\sum_{r=1}^{\infty}\alpha_r(r\lambda)t^{r\lambda-1}}{\sum_{r=1}^{\infty}\alpha_r(1-t^{r\lambda})}, & 0<t<1. \end{cases}
\end{aligned} \tag{9}$$

3. ΕΦΑΡΜΟΓΕΣ

Για την εφαρμογή των αποτελεσμάτων της παρούσας εργασίας που παρουσιάζεται στην ενότητα αυτή, ο αριθμός των τυχαίων μεταβλητών N θεωρείται οτι είναι μία εμφυτευμένη σε Μαρκοβιανή αλυσίδα τυχαία μεταβλητή διωνυμικού τύπου (MVB), βλέπε Koutras and Alexandrou (1995), η κατανομή της οποίας περιγράφει με έναν ενοποιημένο τρόπο κατανομές σχηματισμών (patterns), ροών και συναρτήσεων σάρωσης. Έστω $N \sim MVB(n; A, B)$, $r = 0, 1, \ldots, \ell_n$, όπου n είναι το πλήθος ανεξάρτητων δοκιμών Bernoulli, $\ell_n = \max\{r : \mathsf{P}(N = r) > 0\}$, και A, B είναι κατάλληλοι πίνακες πιθανοτήτων μετάβασης. Ο πίνακας A περιγράφει τις μεταπηδήσεις πρώτης τάξης εντός μιας κατάστασης (within state), και ο πίνακας B τις μεταπηδήσεις πρώτης τάξης μεταξύ καταστάσεων (between states), ενώ ο πίνακας $A + B$ είναι ένας στοχαστικός πίνακας. Η συνάρτηση πιθανογεννήτριας δίνεται από τη σχέση $P_N(z) = \boldsymbol{\pi}_{\mathbf{0}}'(A + zB)^n\mathbf{1}$, με $\boldsymbol{\pi}_{\mathbf{0}}' = (1, 0, \ldots, 0)$ και $\mathbf{1} = (1, 1, \ldots, 1)'$ διανύσματα

γραμμής και στήλης, αντίστοιχα, κατάλληλων διαστάσεων. Επιπλέον,

$$\mathsf{P}(N=r)=\boldsymbol{\pi}_{\mathbf{0}}'\left(\sum_{i=1}^{\binom{n}{r}}D_i^{n,r}\right)\mathbf{1},\quad r=0,1,\ldots,\ell_n, \tag{10}$$

όπου $D_i^{n,r}$ είναι το i-οστό γινόμενο n παραγόντων εκ των A και B με r ακριβώς παράγοντες B, $i=1,\ldots,\binom{n}{r}$. Η παραπάνω έκφραση για την συνάρτηση πιθανότητας της N στην (10) προκύπτει άμεσα από τους Koutras and Alexandrou (1995, Theorem 2.1) με εφαρμογή των σχετικών αναδρομικών σχέσεων.

Για τη συνέχεια η τυχαία μεταβλητή N θεωρείται ότι περιγράφει ροές μήκους τουλάχιστον 2 σε $n=15$ ανεξάρτητες δοκιμές Bernoulli. Σε αυτήν την περίπτωση έχουμε $N\sim MVB(15;A,B)$, $r=0,1,\ldots,5$ ($\ell_{15}=5$), με

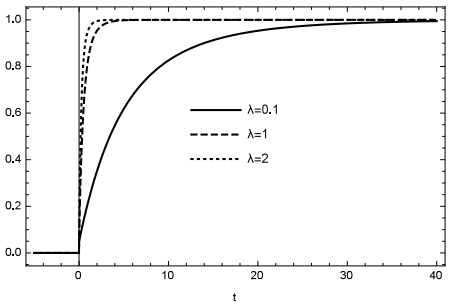
$$A=\begin{bmatrix}q & p & 0\\ q & 0 & 0\\ q & 0 & p\end{bmatrix},\qquad B=\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & p\\ 0 & 0 & 0\end{bmatrix}, \tag{11}$$

όπου p η πιθανότητα επιτυχίας και $q=1-p$. Λαμβάνοντας υπόψη τις (10) και (11), στον Πίνακα 1 δίνονται οι συντελεστές α_r, $r=0,1,\ldots,5$, για διάφορες τιμές του p.

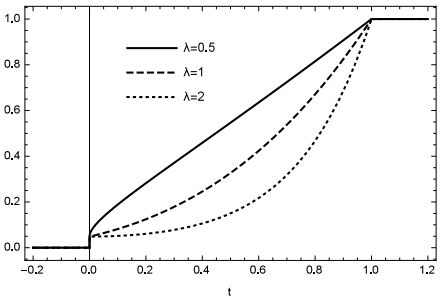
Πίνακας 1: *Οι πιθανότητες* $\mathsf{P}(N=r)$, $r=0,1,\ldots,5$, *με* A *και* B *που δίνονται στην* (11) *για διάφορες τιμές του* p.

	r					
p	0	1	2	3	4	5
0.2	0.610255	0.327552	0.058212	3.8992×10^{-3}	8.1359×10^{-5}	2.4327×10^{-7}
0.5	0.0487366	0.281952	0.439514	0.205505	0.0239563	3.3569×10^{-4}
0.8	6.99646×10^{-5}	0.139445	0.470591	0.337096	5.19047×10^{-2}	8.9335×10^{-4}

Στο Σχήμα 1 δίνονται οι γραφικές παραστάσεις των συναρτήσεων κατανομών των T_1 και T_2 με βάση τις σχέσεις (2), (6) και (3), (7) αντίστοιχα, για $F(t)=1-\exp\{-\lambda t\}$, $t>0$, $\lambda>0$ (εκθετική κατανομή) στην πρώτη περίπτωση και $F(t)=t^{\lambda}$, $0<t<1$, $\lambda>0$ (κατανομή δύναμης) στην δεύτερη περίπτωση. Για τους συντελεστές α_r, $r=0,1,\ldots,5$, χρησιμοποιήθηκαν οι τιμές του Πίνακα 1 για $p=0.5$. Στο Σχήμα 2 δίνονται οι αντίστοιχες γραφικές παραστάσεις των συναρτήσεων κινδύνου των T_1 και T_2 με βάση τις σχέσεις (8) και (9), αντίστοιχα. Σημειώνεται ότι για την $h_T(t)$ με $T=T_1$, προκύπτει ότι $\lim_{t\to\infty}h_T(t)=\lambda$, ενώ για την $h_T(t)$ με $T=T_2$, είναι προφανές ότι $\lim_{t\to0^+}h_T(t)=+\infty$ για $0<\lambda<1$, και $\lim_{t\to1^-}h_T(t)=+\infty$ για $\lambda>0$. Ο έντονος κύκλος που εμφανίζεται στα δύο σχήματα αντιστοιχεί στις τιμές $h_T(0)$, οι οποίες είναι ανεξάρτητες της τιμής της παραμέτρου λ. Τέλος, από το Σχήμα 2α φαίνεται ότι οι αντίστοιχες κατανομές του $T=T_1$ που παράγονται για τις διάφορες τιμές του $\lambda>0$ έχουν φθίνουσα συνάρτηση κινδύνου (decreasing failure rate – DFR) για $t>0$, ενώ από το Σχήμα 2β φαίνεται ότι οι αντίστοιχες

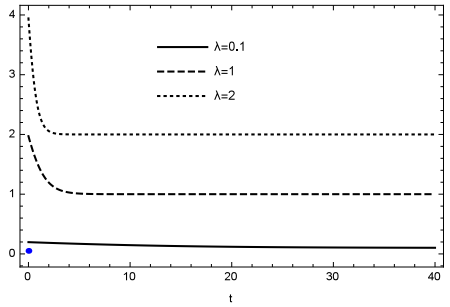


(α)

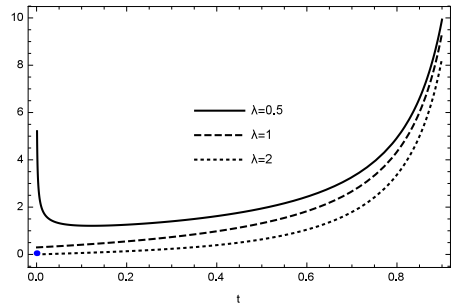


(β)

Σχήμα 1: *Γραφικές παραστάσεις των συναρτήσεων κατανομών των (α)* T_1 *με* $F(t) = 1 - \exp\{-\lambda t\}$, $t > 0$, $\lambda = 0.1, 1, 2$ *και (β)* T_2 *με* $F(t) = t^{\lambda}$, $0 < t < 1$, $\lambda = 0.5, 1, 2$, *και συντελεστές* α_r, $r = 0, 1, \ldots, 5$, *που δίνονται στον Πίνακα 1 για* $p = 0.5$.



(α)



(β)

Σχήμα 2: *Γραφικές παραστάσεις των συναρτήσεων κινδύνου των (α)* T_1 *με* $F(t) = 1 - \exp\{-\lambda t\}$, $t > 0$, $\lambda = 0.1, 1, 2$ *και (β)* T_2 *με* $F(t) = t^{\lambda}$, $0 < t < 1$, $\lambda = 0.5, 1, 2$, *και συντελεστές* α_r, $r = 0, 1, \ldots, 5$, *που δίνονται στον Πίνακα 1 για* $p = 0.5$.

κατανομές του $T = T_2$ που παράγονται για τις διάφορες τιμές του $\lambda \geq 1$ έχουν αύξουσα συνάρτηση κινδύνου (increasing failure rate – IFR) για $0 < t < 1$.

ABSTRACT

The distribution of the minimum and the maximum when the number of continuous and positive random variables of a sample is itself a random variable is studied. Allowing the support of the distribution of the number of the random variables of the sample to contain zero gives rise to problems that have not been systematically studied in the literature. In this work, the moments, the survival functions as well as the hazard rates of the distributions that are derived are also provided. Finally, applications of the results on specific distributions are given.

Ευχαριστίες: Η παρούσα έρευνα έχει χρηματοδοτηθεί από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας μέσω των κονδυλίων της Εθνικής Συμμετοχής 2014–2016 που σχετίζονται με το Ευρωπαϊκό Έργο ISMPH 2014–2016.

ΑΝΑΦΟΡΕΣ

Adamidis, K. and Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statist. Prob Lett*, **39**, 35–42.

Barreto-Souza, W., de Morais, A.L. and Cordeiro, G.M. (2011). The Weibull geometric distribution. *J. Statist. Comput. Simul.*, **81**, 645–657.

Chahkandi, M. and Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate. *Comput. Statist. Data Anal.*, **53**, 4433–4440.

Gibbons, J.D. and Chakraborti, S. (2003). *Nonparametric statistical inference*, 4th ed, New York: Marcel Dekker.

Gupta, R.C., Gupta, P.L. and Gupta, R.D. (1998). Modeling failure time data by Lehmann alternatives. *Commun. Statist. – Theory Meth.*, **27**, 887–904.

Hollander, M. and Wolfe, D.A. (1999). *Nonparametric statistical methods*, 2nd ed, New York: John Wiley & Sons.

Koutras, M.V. and Alexandrou, V.A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach. *Ann. Inst. Statist. Math.*, **47**, 743–766.

Kus, C. (2007). A new lifetime distribution. *Comput. Statist. Data Anal.*, **51**, 4497–4509.

Lehmann, E.L. (1953). The power of rank tests. *Ann. Math. Statist.*, **24**, 23–43.

Louzada, F., Roman, M. and Cancho, V. (2011). The complementary exponential geometric distribution: Model, properties, and a comparison with its counterpart. *Comput. Statist. Data Anal.*, **55**, 2516–2524.

Louzada, F. Bereta, E.M.P. and Franco, M.A.P. (2012). On the distribution of the minimum or maximum of a random number of i.i.d. lifetime random variables. *Applied Math.*, **3**, 350–353.

Morais, A.L. and Barreto-Souza, W. (2011). A compound class of Weibull and power series distributions. *Comput. Statist. Data Anal.*, **55**, 1410–1425.

Tahmasbi, R. and Rezaei, S. (2008). A two-parameter lifetime distribution with decreasing failure rate. *Comput. Statist. Data Anal.*, **52**, 3889–3901.

Tojeiro, C., Louzada, F., Roman, M. and Borges, P. (2014). The complementary Weibull geometric distribution. *J. Statist. Comput. Simul.*, **84**, 1345–1362.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 30ου Πανελληνίου Συνεδρίου Στατιστικής (2017), σελ.44-56

ΤΖΑΚΠΟΤ ΚΑΙ ΜΑΘΗΜΑΤΙΚΑ

Μωυσιάδης Πολυχρόνης
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
cmoi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Η λέξη Τζάκποτ ήταν μέχρι πρότινος γνωστή μόνο στους παίκτες των Καζίνο. Τα τελευταία χρόνια όμως, ιδιαίτερα μετά την καθιέρωση των τυχερών παιχνιδιών με αριθμούς, όπως ΛΟΤΤΟ, ΤΖΟΚΕΡ και άλλα, έχει ενταχθεί στην καθημερινότητα και απασχολεί σε μεγάλο βαθμό τους πολίτες, οι οποίοι κατά καιρούς διατυπώνουν διάφορα ερωτήματα όπως γιατί συμβαίνουν πολλά διαδοχικά Τζάκποτ, γιατί τα Τζάκποτ συμβαίνουν κυρίως στις μεγάλες γιορτές, μήπως δεν είναι έντιμες οι κληρώσεις και διάφορα σχετικά.

Στην εργασία αυτή προσπάθησα να δώσω μια εξήγηση σε τέτοια ερωτήματα, χρησιμοποιώντας βασικές αρχές της Στατιστικής και λαμβάνοντας δεδομένα από τον ιστοχώρο του ΤΖΟΚΕΡ, που έχει αναρτημένα όλα τα αποτελέσματα και είναι προσβάσιμος στον κάθε ενδιαφερόμενο.

Λέξεις Κλειδιά: τζόκερ, τζάκποτ, κατανομή, ροές.

1. ΜΗΧΑΝΗΜΑΤΑ ΤΥΧΕΡΩΝ ΠΑΙΧΝΙΔΙΩΝ

Τα τυχερά παιχνίδια και οι Λοταρίες έχουν πολλούς οπαδούς αλλά και επικριτές. Στηρίζονται στην επιθυμία των ανθρώπων να κερδίσουν μεγάλα χρηματικά ποσά διακινδυνεύοντας στη θεά τύχη πολύ μικρότερα. Είναι μία πηγή εκατοντάδων εκατομμυρίων επαναλήψεων τυχαίων φαινομένων χρήσιμη για τους ελέγχους των νόμων των πιθανοτήτων.

Η Ελλάδα είναι μία από τις πρώτες χώρες στον κόσμο στα κατά κεφαλήν έξοδα για τυχερά παιχνίδια, λοταρίες κλπ.

Το τζάκποτ είναι μια έξυπνη επινόηση που αύξησε τους παίκτες που συμμετέχουν σ' αυτά τα παιχνίδια και κατά συνέπεια την κερδοφορία των διοργανωτών τυχερών παιχνιδιών. Εφαρμόστηκε αρχικά στα Μηχανήματα Τυχερών Παιχνιδιών (ΜΤΠ), δηλ. στα φρουτάκια ή κουλοχέρηδες, κλπ, των Καζίνο. Κάποια από αυτά τα ΜΤΠ συμμετέχουν σε ομάδες ΜΤΠ είτε του ίδιου του Καζίνο στο οποίο είναι

τοποθετημένα είτε με άλλα ΜΤΠ άλλων Καζίνο, σε αυτό που αποκαλείται προοδευτικό Τζάκποτ (Progressive Jackpot). Αυτό λειτουργεί ως εξής: Κάθε φορά που κάποιος παίκτης παίζει σε ένα από τα ΜΤΠ που συμμετέχουν, και δεν κερδίζει, ένα μικρό ποσοστό του ποσού που διακυβεύτηκε προστίθεται στο προοδευτικό Τζάκποτ. Έτσι το προοδευτικό Τζάκποτ αυξάνεται συνεχώς, μέχρι να βρεθεί ο τυχερός νικητής, που σε κάποιο από τα ΜΤΠ της συγκεκριμένης ομάδας θα πετύχει κάποιον εξαιρετικά σπάνιο συνδυασμό.

Στα ΜΤΠ που συμμετέχουν στα προοδευτικά τζάκποτ υπάρχει σε εμφανές σημείο η ένδειξη με το ποσό του Τζάκποτ που αυξάνει, ώστε οι παίκτες να ελπίζουν ότι θα κερδίσουν το μεγάλο αυτό ποσό, οπότε παρακινούνται να παίξουν περισσότερο. Σε ΜΤΠ που παίζονται στο ίντερνετ και συμμετέχουν σε ομάδες, το προοδευτικό Τζάκποτ φαίνεται στους τίτλους των παιχνιδιών.

Στον κανονισμό που διέπει τη λειτουργία των τυχερών παιχνιδιών στην Ελλάδα (Τ/6736, ΦΕΚ 929 Β) του 2003 και στο άρθρο 19 εδ. 30 καθορίζεται ότι το προοδευτικό τζάκποτ πρέπει να συμπεριλαμβάνεται στο ποσοστό απόδοσης των μηχανημάτων και επομένως συνυπολογίζεται από τις κατασκευαστικές εταιρείες τέτοιων ΜΤΠ.

Το 1986 η αμερικάνικη εταιρεία IGT κατασκεύασε και προώθησε τα πρώτα jackpot slots, δηλαδή μηχανήματα σχεδιασμένα να συμμετέχουν σε ομάδες προοδευτικού τζάκποτ. Τα μηχανήματα είχαν συνδεθεί μεταξύ τους με τηλεφωνικές γραμμές μέσω μόντεμ. Η IGT χρηματοδότησε το πρώτο εκατομμύριο δολάρια για να ξεκινήσει το τζάκποτ, αλλά από εκεί και πέρα, η αύξηση του τζάκποτ εξαρτιόταν από το πόσοι άνθρωποι θα έπαιζαν στα συγκεκριμένα ΜΤΠ. Αυτό αύξησε τους παίκτες που έπαιζαν και το πρώτο προοδευτικό τζάκποτ ανήλθε στο ποσό των $4,988,842.17. Το ποσό αυτό κερδήθηκε από έναν παίκτη στο Reno των Ηνωμένων Πολιτειών την 1η Φεβρουαρίου του 1987.

Το Μάιο του 2009, ένας Έλληνας παίζοντας στο Mega Moolah στο καζίνο River Belle, κέρδισε 6.3 εκατομμύρια ευρώ από προοδευτικό Τζάκποτ. Τέτοιες ειδήσεις αυξάνουν το ενδιαφέρον για τα μηχανήματα τυχερών παιχνιδιών και το προοδευτικό τζάκποτ φαίνεται να είναι ο καλύτερος διαφημιστής αυτών των παιχνιδιών.

2. ΤΖΟΚΕΡ

2.1 Το παιχνίδι

Το Τζόκερ είναι παιχνίδι αριθμών και άρχισε να παίζεται στην Ελλάδα από το έτος 1997. Οι κανόνες του παιχνιδιού υπάρχουν στον αντίστοιχο ιστοχώρο και όπως θα αποδειχθεί στη συνέχεια υπάρχει πολύ μεγάλη πιθανότητα να μην υπάρξει νικητής.

Στο παιχνίδι του Τζόκερ υπάρχουν 8 κατηγορίες νικητών, που φαίνονται στον Πίνακα 1, ο οποίος υπάρχει στον ιστοχώρο του. Από τον πίνακα αυτόν διαπιστώνουμε ότι τα κέρδη των έξι τελευταίων κατηγοριών είναι καθορισμένα. Στις δύο πρώτες κατηγορίες τα κέρδη καθορίζονται από τη συμμετοχή των παικτών.

Πίνακας 1. *Κατηγορίες νικητών του τζόκερ*

ΚΑΤΗΓΟΡΙΕΣ ΕΠΙΤΥΧΙΩΝ	I	II	III	IV	V	VI	VII	VIII
ΣΩΣΤΕΣ ΠΡΟΒΛΕΨΕΙΣ	5+1	5	4+1	4	3+1	3	2+1	1+1
ΚΕΡΔΟΣ ΑΝΑ ΕΠΙΤΥΧΙΑ	*	*	2.500€	50€	50€	2€	2€	1,50€

Μάλιστα, αν στην κλήρωση δεν υπάρξει νικητής σε μία από τις κατηγορίες αυτές, το ποσό μεταφέρεται στην επόμενη κλήρωση στην κατηγορία του, και η συγκεκριμένη κλήρωση χαρακτηρίζεται ως ΤΖΑΚΠΟΤ. Προφανώς αυτό το τζάκποτ είναι διαφορετικό από το προοδευτικό τζάκποτ των καζίνο, όμως είναι στην ίδια λογική της αύξησης των κερδών για άγρα παικτών και πράγματι το διαπιστώνουμε αυτό κάθε φορά που έχει μεγάλο αριθμό διαδοχικών τζάκποτ.

Οι πιθανότητες νίκης στις 2 πρώτες κατηγορίες είναι:

Για την κατηγορία I υπάρχουν $\binom{45}{5}\cdot\binom{20}{1}=24435180$ δυνατές στήλες και άρα έχουμε πιθανότητα $\frac{1}{24435180}=0.0000000409$.

(δηλ. 4 εκατοντάκις εκατομμυριοστά, ή αλλιώς η πιθανότητα να γίνει θαύμα)

Για την κατηγορία II υπάρχουν $\binom{45}{5}\cdot\binom{20}{0}=1221759$ δυνατές στήλες και άρα έχουμε πιθανότητα $\frac{1}{1221759}=0.000000814$.

(δηλ. 8 δεκάκις εκατομμυριοστά, επίσης εξαιρετικά μικρή πιθανότητα).

Γιατί όμως συμβαίνουν ΤΖΑΚΠΟΤ; ή αλλιώς γιατί δεν βρίσκεται νικητής στην κατηγορία I (ή και στη II) αφού παίζονται τόσα εκατομμύρια στήλες; Η απάντηση είναι φανερή από τους παραπάνω υπολογισμούς. Αναμένουμε μία στήλη στις 24.4 εκατομμύρια στήλες να κερδίσει, και όπως βλέπουμε στον Πίνακα 2 οι στήλες που παίζονται είναι πολύ λιγότερες και κυμαίνονται γύρω στα 3 με 4 εκατομμύρια. Όσο τα διαδοχικά τζάκποτ αυξάνονται, τόσο και οι στήλες που παίζονται αυξάνονται. Μάλιστα, όταν τα διαδοχικά τζάκποτ είναι περισσότερα από 7-8, οι στήλες που παίζονται διπλασιάζονται, τριπλασιάζονται και στην προκειμένη περίπτωση που έφτασαν τα 16 τζάκποτ, οι στήλες που παίχτηκαν ήταν πέντε-έξι φορές περισσότερες από τις συνήθεις.

Πίνακας 2. *Απόσπασμα αποτελεσμάτων του τζόκερ*

ΑΡΧΕΙΟ ΚΛΗΡΩΣΕΩΝ ΤΖΟΚΕΡ									ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΙΑΛΟΓΗΣ				
ΑΑ Γενικό	ΑΑ Κληρ.	Ημερομηνία	Αριθμοί					Τζόκερ	Στήλες	(5+1) Επιτυχίες	(5+1) Κέρδη/Επιτ.	(5) Επιτυχίες	(5) Κέρδη/Επιτ.
2003	1784	16/2/2017	9	11	33	43	44	13	21.854.909	1	16.428.350	8	120.885
2002	1783	12/2/2017	4	21	33	34	36	16	20.143.417	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	6	148.558
2001	1782	9/2/2017	11	17	29	33	43	16	17.350.176	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	8	95.968
2000	1781	5/2/2017	7	10	13	14	30	17	16.459.650	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	7	104.049
1999	1780	2/2/2017	1	8	13	30	42	3	13.271.199	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	5	117.450
1998	1779	29/1/2017	14	23	31	41	45	7	11.816.087	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	8	65.358
1997	1778	26/1/2017	8	12	14	23	45	16	9.691.135	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	1	428.833
1996	1777	22/1/2017	11	14	18	27	45	19	9.029.050	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	3	133.178
1995	1776	19/1/2017	20	22	29	31	32	5	6.940.075	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	3	102.366
1994	1775	15/1/2017	8	22	27	29	38	4	6.692.612	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	14	21.153
1993	1774	12/1/2017	20	21	25	30	38	11	5.284.868	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	5	46.771
1992	1773	8/1/2017	11	22	26	35	36	18	4.888.462	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	1	216.314
1991	1772	5/1/2017	4	7	25	37	42	13	5.128.947	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	4	56.739
1990	1771	1/1/2017	14	16	17	26	34	10	5.635.365	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	2	222.162
1989	1770	29/12/2016	5	26	30	31	41	17	4.405.871	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ
1988	1769	24/12/2016	1	3	31	35	42	7	3.053.461	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	1	135.116
1987	1768	22/12/2016	12	13	20	34	39	11	3.463.760	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	1	153.271
1986	1767	18/12/2016	5	9	11	28	31	19	5.047.655	1	2.500.000	12	18.613
1985	1766	15/12/2016	1	14	27	39	41	2	4.450.985	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	14	14.068
1984	1765	11/12/2016	10	14	33	34	42	6	4.494.881	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	4	49.725
1983	1764	8/12/2016	8	15	22	40	45	6	3.883.157	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	2	85.915
1982	1763	4/12/2016	12	17	28	30	36	8	3.846.171	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	2	85.097
1981	1762	1/12/2016	14	29	35	43	45	1	3.337.128	ΤΖΑΚΠΟΤ	ΤΖΑΚΠΟΤ	1	147.668
1980	1761	27/11/2016	2	31	33	35	43	20	3.716.423	1	600.000	2	82.226
1979	1760	24/11/2016	11	13	25	38	45	11	4.542.766	1	1.667.979	2	100.509

Σε τέτοιες περιπτώσεις ακόμη και οι εφημερίδες διαφημίζουν και προτρέπουν τους πολίτες να παίξουν με δημοσιεύματα όπως:

- «Δύο υπερτυχερούς ανέδειξε η κλήρωση του Τζόκερ (6-11-2014) για τα 18,4 εκατ. ευρώ, το δεύτερο μεγαλύτερο ποσό στην ιστορία του παιχνιδιού με τα δελτία που κατατέθηκαν να ξεπερνούν τα 5 εκατομμύρια (και στήλες πάνω από 39 εκατ.).»
- «Στις 16-4-2010, τρεις τυχεροί μοιράστηκαν το τζακποτ των 19,2 εκ. Ευρώ».

Υπάρχουν όμως και αρνητικά δημοσιεύματα, όπως:

- «Να παρέμβει ο οικονομικός εισαγγελέας και να μάθει πως γίνεται κάθε γιορτινή περίοδο να γίνεται ΤΖΑΚ ΠΟΤ και ένας άγνωστος να μαζεύει όλο το χρήμα.»
- «Η απάτη του Τζόκερ και το σανό που ταΐζουν στον Έλληνα.»

ή κυκλοφορούν βίντεο που ενημερώνουν για την «απάτη».

Συμβαίνει πράγματι κάτι επιλήψιμο;

Αρκετοί από μας τους στατιστικούς ενοχληθήκαμε από ραδιόφωνα ή τηλεοπτικά μέσα, που σε περιόδους που εμφανίζονται πολλά διαδοχικά τζακποτ θέλουν να πληροφορήσουν τους ακροατές ή τους θεατές τους για την πιθανότητα να συμβούν αυτά τα τζακποτ. Θα προσπαθήσω, στη συνέχεια, να δώσω τη δική μου προσέγγιση σ' αυτό το ερώτημα, δηλαδή να εξετάσω κατά πόσον ήταν δυνατόν να συμβούν τα τζακποτ έτσι όπως συνέβησαν.

2.2 Ένα κατασκευασμένο πρόβλημα ανάλογο του Τζόκερ αλλά απλούστερο

Πριν ασχοληθούμε με τα αποτελέσματα του Τζόκερ, ας μελετήσουμε πρώτα ένα κατασκευασμένο παιχνίδι που έχει τη λογική του τζόκερ, δηλαδή της ύπαρξης τζάκποτ και της μεταφοράς των κερδών σε επόμενες κληρώσεις.

Ας θεωρήσουμε ότι 4 άτομα παίζουν ένα τυχερό παιχνίδι ρίχνοντας ο καθένας ένα κανονικό ζάρι, στο οποίο ισχύουν οι παρακάτω κανόνες:

- Πριν το παιχνίδι, ποντάρουν ο καθένας από ένα σταθερό ποσό, π.χ. 1€.
- Αν ένας ή περισσότεροι φέρουν 6-ρι, τότε έχουμε νικητή ή νικητές που μοιράζονται το ποσόν που συγκεντρώθηκε.
- Αν κανείς δεν φέρει 6-ρι, τότε έχουμε τζακποτ και το ποσό μεταφέρεται στο επόμενο παιχνίδι.

Πόσα τζακποτ μπορεί να συμβούν διαδοχικά μέχρι να βρεθεί νικητής;

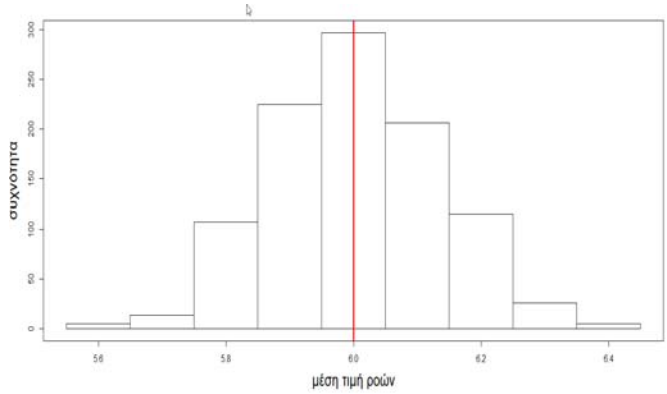
Σε κάθε παιχνίδι θεωρούμε ότι έχουμε 4 ανεξάρτητες ρίψεις ενός ζαριού. Επομένως μετά από m γύρους παιχνιδιού, θα έχω $n = 4 \cdot m$ ανεξάρτητες ρίψεις του ζαριού. Για να μελετήσω τη συμπεριφορά του παιχνιδιού, έκανα προσομοίωση εκτελώντας $n = 10.000$ ρίψεις (δηλαδή 2500 γύρους του παραπάνω παιχνιδιού) και υπολόγισα το διάνυσμα των ροών αποτυχιών, το διάνυσμα των τζακποτ, και την κατανομή των τζακποτ.

Επίσης υπολόγισα τη μέση τιμή και το μέγιστο των ροών αποτυχιών.

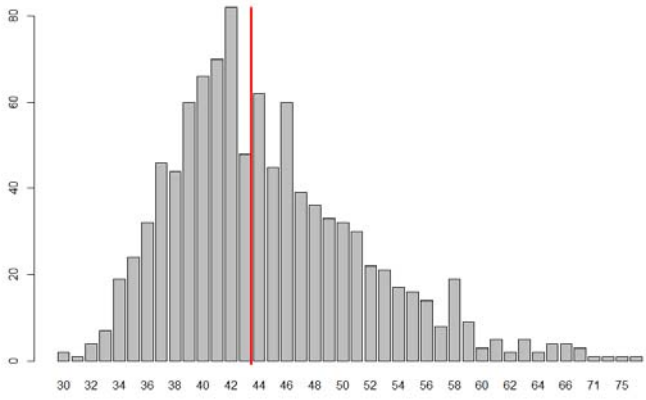
Για τις δύο αυτές τιμές βρήκα την κατανομή τους επαναλαμβάνοντας 1000 φορές την προσομοίωση των «10000 ρίψεων».

Οι μέσες τιμές των ροών ακολουθούν μάλλον κανονική κατανομή γύρω από το 6, όπως αναμενόταν (Σχ.1), όμως το μέγιστο των ροών ήταν σχεδόν πάντα αρκετά μεγάλο στις 10000 ρίψεις (Σχ.2). Και τα διαδοχικά τζάκποτ σε 10.000 ρίψεις, φτάνουν ακόμη και τα 15 (Στο Σχ.3 ήταν 11). Αυτό το τελευταίο είναι εξαιρετικά ενδιαφέρον και δείχνει ότι και σε αυτό το απλό παιχνίδι μπορούν να συμβούν πολλά διαδοχικά τζάκποτ.

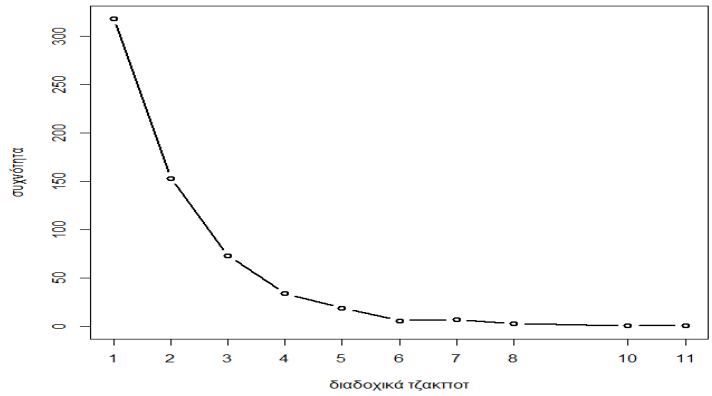
Σχήμα 1. *Ιστόγραμμα ροών 1000 επαναλήψεων 10000 ρίψεων*



Σχήμα 2. *Ραβδόγραμμα μέγιστου ροών 10000 ρίψεων σε 1000 επαναλήψεις*



Σχήμα 3. *Συχνότητα εμφάνισης διαδοχικών τζακποτ*



2.3 Η πρώτη κατηγορία κερδών του Τζόκερ

Τα αποτελέσματα των κληρώσεων του Τζόκερ υπάρχουν διαθέσιμα στον ιστοχώρο του Τζόκερ (http://www.opap.gr/el/web/guest/joker-draw-results), και μπορούμε να τα κατεβάσουμε κάνοντας κλικ στο κουμπί «Αρχείο Αποτελεσμάτων». Για τα πρώτα χρόνια 1997-2002 υπάρχουν μόνο οι αριθμοί που κληρώθηκαν χωρίς άλλα στοιχεία. Από το έτος 2003 έως σήμερα είναι ανηρτημένα όλα τα στοιχεία όπως: στήλες που παίχτηκαν, αριθμοί που κληρώθηκαν, κέρδη που διανεμήθηκαν, πότε και σε ποια κατηγορία έγινε τζάκποτ κλπ.

Τμήμα των αποτελεσμάτων με τρεις πρόσφατες κληρώσεις δίνεται στον Πίνακα 3, όπου μας ενδιαφέρουν κυρίως αυτά που έχω κυκλώσει.

Πίνακας 3. *Απόσπασμα πλήρους πίνακα αποτελεσμάτων τριών κληρώσεων*

ΑΠΟΤΕΛΕΣΜ

ΚΛΗΡ.	ΗΜ/ΝΙΑ	ΚΛΗΡΩΘΕΝΤΕΣ ΑΡΙΘΜΟΙ 5 ΑΠΟ 45					TZOKEP 1 από 20	ΣΥΝ. ΣΤΗΛΩΝ	5 + 1		5	
									ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ	ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ
1793	19/3/2017	17	37	38	29	31	15	6.011.833	1	3.095.659,02	5	99.212,95
1792	16/3/2017	20	35	4	14	27	10	5.198.670	-	ΤΖΑΚ-ΠΟΤ	-	ΤΖΑΚ-ΠΟΤ
1791	12/3/2017	41	8	10	38	28	6	5.125.238	-	ΤΖΑΚ-ΠΟΤ	3	75.597,26

ATA TZOKEP 2017

4 + 1		4		3 + 1		3		2 + 1		1 + 1	
ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ	ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ	ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ	ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ	ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ	ΕΠ/ΧΙΕΣ	ΚΕΡΔΗ
50	2.500,00	1.206	50,00	1.759	50,00	43.233	2,00	18.921	2,00	84.922	1,50
51	2.500,00	692	50,00	1.571	50,00	29.188	2,00	21.063	2,00	99.694	1,50
30	2.500,00	857	50,00	1.442	50,00	32.399	2,00	18.099	2,00	88.573	1,50

Για τη συνέχεια, έκανα την ακόλουθη υπόθεση:

> *Θεωρώ ότι οι στήλες κάθε κλήρωσης, που συνήθως είναι αρκετά εκατομμύρια, είναι «ανεξάρτητες» επαναλήψεις του τυχαίου φαινομένου* ***«επιλέγουμε 5 αριθμούς από το 1 έως το 45 & 1 αριθμό από το 1 έως το 20»,*** *στο οποίο αναζητείται μια άγνωστη εξάδα (5+1) αριθμών, που γίνεται γνωστή με την κλήρωση.*

Προφανώς αυτό δεν είναι ακριβές, διότι λόγω των συστημάτων που παίζονται επιλέγονται συστηματικές στήλες αλλά και επειδή παίζουν πολλές στήλες οι ίδιοι άνθρωποι. Όμως λόγω του πολύ μεγάλου πλήθους, νομίζω ότι η υπόθεση ανεξαρτησίας δημιουργεί πολύ μικρό σφάλμα και επιτρέπει τη μελέτη.

Εάν σε μία κλήρωση παίχτηκαν n στήλες και είχαμε k επιτυχίες στην κατηγορία I, θεωρήθηκαν k τυχαία επιλεγμένοι αριθμοί στο διάστημα 1 έως n, ως

σημεία στα οποία υπήρξε νικητής, δηλαδή ήταν «τυχερές στήλες», ενώ όλες οι άλλες στήλες ήταν οι άτυχες στήλες.

Αθροίστηκαν οι διαδοχικές άτυχες στήλες και σχημάτισαν ροές άτυχων στηλών μεταξύ των τυχερών στηλών.

Η καταγραφή των ροών έγινε ως εξής. Έστω ότι σε μια κλήρωση π.χ. στην n όπου είχαν παιχτεί $cols(n)$ στήλες είχαμε μόνο μία επιτυχία μετά από k τζάκποτ ($k = 0,1,2,...$). Θεωρούμε ότι η επιτυχία έγινε σε μία τυχαία από τις στήλες που παίχτηκαν, δηλαδή στη στήλη $sample(n) = [r \cdot cols(n)]$, όπου r τυχαίος αριθμός στο $(0,1)$ και $[a]$ είναι το ακέραιο μέρος του a. Η διαφορά του αριθμού $sample(n)$ από τις στήλες που παίχτηκαν $cols(n)$ μας δίνει τις άτυχες που θα πρέπει να προσμετρηθούν στις στήλες που ακολουθούν μέχρι την επόμενη επιτυχία και συμβολίζεται:

$$resd(n) = cols(n) - sample(n). \tag{1}$$

Αθροίζοντας τις στήλες που προηγήθηκαν της τυχερής στήλης, αυτές που παίχτηκαν στα προηγούμενα k τζάκποτ και τις υπόλοιπες από την προηγούμενη επιτυχία βρίσκουμε τον αριθμό των ροών δηλ. των άτυχων στηλών που τις συμβολίζουμε με το del51. Έχουμε λοιπόν:

$$del51(n) = sample(n) - 1 + \sum_{i=1}^{k} cols(n-i) + resd(n-k-1) \tag{2}$$

Η καταγραφή έγινε με το Excel και ο Πίνακας 4 περιέχει ένα απόσπασμα των υπολογισμών.

Ας θεωρήσουμε ότι η γραμμή με το βέλος στον πίνακα 4 είναι η κλήρωση με αριθμό n στην οποία παίχτηκαν $cols(n) = 6374720$ στήλες. Τότε ο κυκλωμένος με έλλειψη αριθμός στη στήλη sample είναι ο $sample(n) = 3446248$, που προέκυψε με μια τυχαία γεννήτρια στο διάστημα $(1, cols(n))$. Ο αριθμός στο τριγωνικό πλαίσιο είναι ο $resd(n) = 2928472$ και είναι η διαφορά του τυχαίου από το σύνολο στηλών, όπως περιγράφεται στη σχέση (1). Ο αριθμός στο τετραγωνικό πλαίσιο είναι ο ζητούμενος $del51(n) = 33593855$ που προέκυψε σύμφωνα με τη σχέση (2) από το άθροισμα των οκτώ αριθμών που είναι μέσα στα ελλειπτικά πλαίσια.

Παρατηρούμε, επίσης, στον πίνακα 4 ότι επτά κληρώσεις πριν την σημειωμένη είχαμε 2 επιτυχίες στην κατηγορία Ι. Τι κάνουμε στην περίπτωση αυτή;

Πίνακας 4. *Καταγραφή ροών διαδοχικών κληρώσεων στο Excel*

cols	scs51	del51	resd	sample	win51
19152860	0				0,00
20264479	0				0,00
22169174	1	150763410	6278321	15890853	15500000,00
3679104	0				0,00
4266116	1	10015930	4207611	58505	790549.39
4292622	2	5951806		1744195	300000,00
		762157	1786270	2506352	
3814331	0				0,00
5175044	0				0,00
3967301	0				0,00
4478738	0				0,00
5542190	0				0,00
5383733	0				0,00
6374720	1	33593855	2928472	3446248	3456237.67
3424590	0				0,00
3794674	0				0,00

Συμβολίζω με m την κλήρωση αυτή, και προφανώς πρέπει να επιλέξω δύο τυχαίους αριθμούς στο διάστημα $(1, cols(m)) = (1, 4292622)$, που εδώ έτυχε να είναι οι $sample(m_1) = 1744195$ και $sample(m_2) = 2506352$, τους οποίους έχω διατάξει σε αύξουσα σειρά. Κατά συνέπειαν έχω και δύο τιμές στη στήλη del51, εκ των οποίων η πρώτη βγαίνει όπως προηγούμενα, ενώ η δεύτερη ως εξής:

$$del51(m_2) = sample(m_2) - sample(m_1) \qquad (3)$$

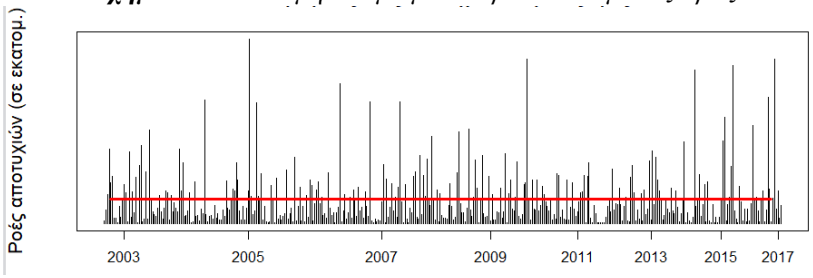
Αν υπάρχουν περισσότερες επιτυχίες εργαζόμαστε ανάλογα. Σημειώνω ότι υπήρξε κλήρωση με 10 επιτυχίες, οι οποίες φυσικά μοιράστηκαν το έπαθλο.

Η καταγραφή έγινε για 1575 κληρώσεις. Αξιοποιώντας τις δυνατότητες του Excel άφησα ως συνάρτηση τον τυχαίο αριθμό της στήλης sample εκεί που είχαμε μία επιτυχία, ώστε με κάθε προσθήκη έστω και ενός χαρακτήρα να επιλέγεται νέος τυχαίος αριθμός. Αυτό μετατρέπει την όλη διαδικασία σε μια δυναμική διαδικασία η οποία δεν ισχύει για μία μόνο στατική επιλογή των τυχαίων αριθμών στις κληρώσεις με επιτυχία, αλλά κάθε φορά μεταβάλλεται.

Η κατανομή εκτιμώμενων ροών και η θέση του μέσου όρου (23.3 εκατομμύρια) φαίνεται στο Σχήμα 4.

Αξίζει να σημειωθεί ότι η κατανομή των ροών δεν διαφοροποιείται σημαντικά, όποιο και από τα διαφορετικά διανύσματα del51, που προκύπτουν από διαφορετικές επιλογές τυχαίων αριθμών, πάρουμε, για τη γραφική παράσταση του Σχήματος 4. Οι

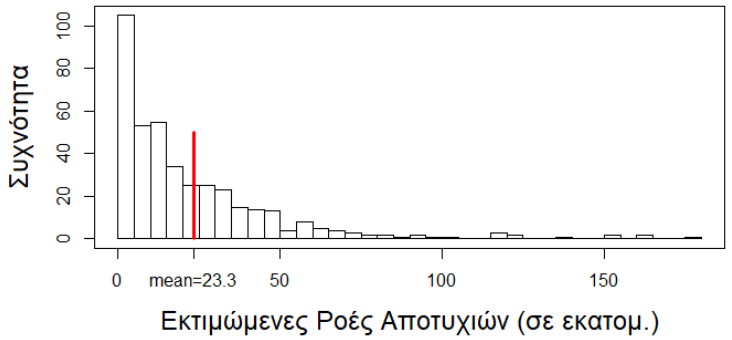
Σχήμα 4. *Κατανομή εκτιμώμενων ροών και μέσος όρος*



θέσεις των μεγίστων κάπως αλλάζουν, αλλά ο μέσος όρος είναι γύρω από το 23.3 εκατομμύρια. Βέβαια από τον υπολογισμό των πιθανοτήτων περιμέναμε να είναι γύρω στα 24.4 εκ. αλλά πιστεύω ότι αυτό οφείλεται στο ότι στην πράξη δεν είναι πράγματι ανεξάρτητες οι στήλες και η συστηματική επιλογή στηλών από διάφορα έξυπνα συστήματα, συντομεύει την επίτευξη επιτυχίας.

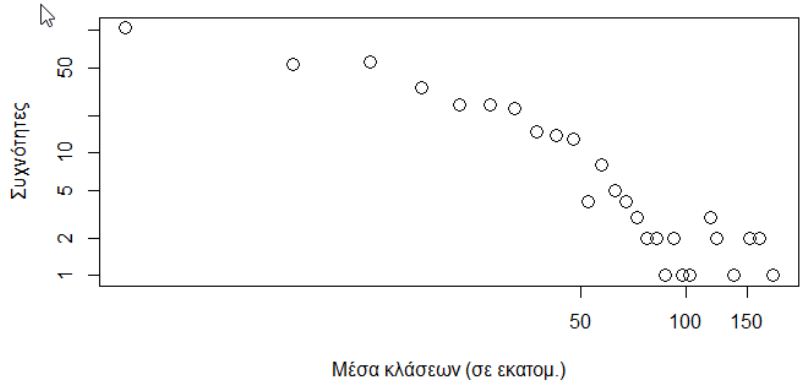
Από το ιστόγραμμα του Σχήματος 5, παρατηρούμε ότι οι ροές ακολουθούν μια εκθετική μείωση, όμως φτάνουν και μέχρι πολύ μεγάλα ποσά πάνω από 150 εκατομμύρια στήλες χωρίς επιτυχία, με μέσο όρο όμως περίπου σταθερό.

Σχήμα 5. Ιστόγραμμα εκτιμώμενων ροών αποτυχιών από 2003-2017



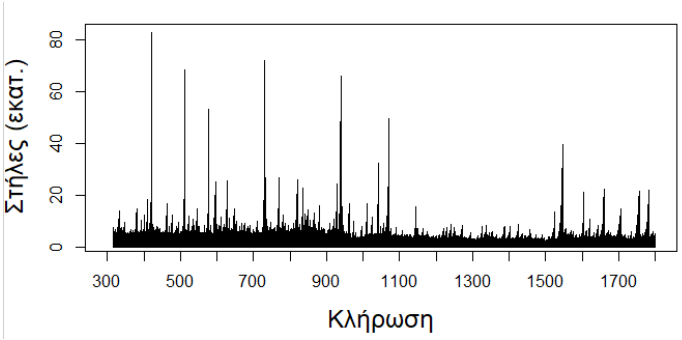
Οι ροές σε Λογο-λογαριθμική γραφική παράσταση αποκαλύπτουν δυναμοκατανομή, κάτι που φαίνεται και από το Σχήμα 5. Αυτό ενισχύεται με την εκτέλεση της εντολής power.law.fit από το πακέτο poweRlaw της R, που δίνει πιθανότητα p=0,94 και άρα δεν απορρίπτεται η υπόθεση ότι ακολουθεί Power Law κατανομή. (Σχ.6)

Σχήμα 6. Λογο-λογαριθμική παράσταση ροών αποτυχιών από 2003-2017



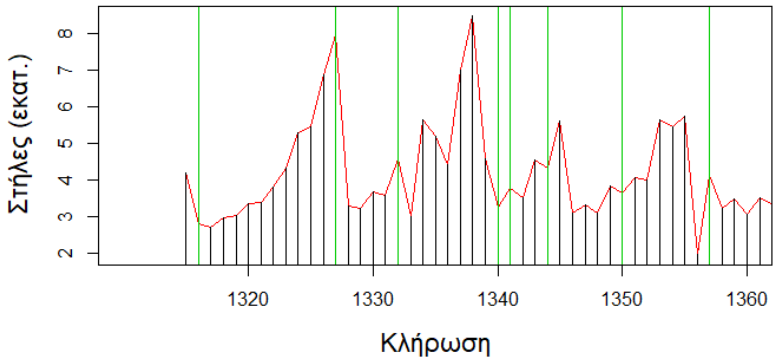
Στο Σχήμα 7 δίνουμε τη γραφική παράσταση των στηλών που παίχτηκαν όλα αυτά τα χρόνια. Παρατηρούμε ότι υπάρχει μια εμφανής μείωση των στηλών που παίχτηκαν κατά το δεύτερο μισό του διαστήματος, που αντιστοιχεί στα έτη μετά το 2010, δηλαδή μετά την κρίση.

Σχήμα 7. Διάγραμμα στηλών που παίχτηκαν



Στο Σχήμα 8 επέλεξα 50 κληρώσεις, για να φανεί πως μεταβάλλονται οι στήλες που παίζονται όταν αυξάνεται ο αριθμός των τζακποτ, και επομένως και του ποσού που θα διανεμηθεί. Παρατηρούμε ότι στην αρχή μετά από μια επιτυχία παίζονται λίγες στήλες, από τους σταθερούς παίκτες. Όταν τα τζάκποτ αυξάνονται τότε οι στήλες πολλαπλασιάζονται, προφανώς και από έκτακτους παίκτες που αναζητούν το μεγάλο κέρδος.

Σχήμα 8. Αύξηση στηλών λόγω τζακποτ για 50 κληρώσεις



2.3 Η δεύτερη κατηγορία κερδών του Τζόκερ

Για την κατηγορία ΙΙ έγινε ανάλογη δουλειά, μόνο που εδώ είναι πιο δύσκολο να υπολογιστούν οι ροές στην κατηγορία αυτή. Ο λόγος είναι ότι στην κατηγορία αυτή έχουμε πολύ λιγότερα τζάκποτ, αλλά πάρα πολλές κληρώσεις με πολλές επιτυχίες. Στην κλήρωση της 21/11/2004 που είχαμε μέγιστο 10 στις επιτυχίες της κατηγορίας Ι, είχαμε και μέγιστο στις επιτυχίες της κατηγορίας ΙΙ που ήταν 149.

Κατέγραψα τις ροές αποτυχιών της κατηγορίας ΙΙ μόνο για τα δύο πρώτα χρόνια 2003 και 2004 και προέκυψαν ανάλογα σχήματα, που παραλείπονται. Έτσι η κατανομή εκτιμώμενων ροών είναι όπως στο σχήμα 4, μόνο που η θέση του μέσου όρου είναι στα 1.27 εκατομμύρια. Ομοίως και το ιστόγραμμα εκτιμώμενων ροών για την κατηγορία ΙΙ από 2003-2004 είναι παρόμοιο με το Σχήμα 5, με διαφορετικό μέσο όρο. Εδώ ο μέσος αναμένονταν να είναι 1.22 και προέκυψε λίγο μεγαλύτερος.

3. ΕΠΙΛΟΓΟΣ

Από την ανωτέρω ανάλυση που έκανα, παρατηρώ ότι είναι αναμενόμενο να υπάρξουν πολλά διαδοχικά τζάκποτ τόσο στο κανονικό παιχνίδι του Τζόκερ, όσο και στο απλούστερο που πρότεινα στην παράγραφο 2.2. Στο απλούστερο παιχνίδι υπάρχει η δυνατότητα να γίνουν πολλές επαναλήψεις σειράς κληρώσεων και έτσι να μελετηθεί ακόμη περισσότερο η συμπεριφορά των τζάκποτ. Στο κανονικό παιχνίδι, όμως, δεν είναι τόσο εύκολη η προσομοίωση, αφού είναι δύσκολη η περιγραφή της συμπεριφοράς των παικτών μετά από κάποια επιτυχή κλήρωση που σταματά τα τζάκποτ. Συνήθως μειώνεται το ενδιαφέρον των παικτών για τις πρώτες κληρώσεις μετά από αυτήν την επιτυχία και στη συνέχεια όταν και αν αυξηθούν τα τζάκποτ, τότε αναζωπυρώνεται και το ενδιαφέρον των παικτών, όχι βέβαια πάντα, αφού υπάρχουν και μειώσεις στηλών μετά από πολλά τζάκποτ, όπως φαίνεται στο σχήμα 8.

Κατά τη γνώμη μου, τα αποτελέσματα που εμφανίστηκαν δεν απέχουν πολύ από τα αναμενόμενα, τουλάχιστον ως προς το πλήθος των τζάκποτ και με τις ροές των αποτυχιών. Σχετικά με τα ποσά που διανέμονται, αυτό είναι άλλο θέμα με το οποίο δεν ασχολήθηκα. Σχετικά με το χρόνο που συμβαίνουν τα τζάκποτ και οι επιτυχίες, αυτό ίσως οφείλεται στο γεγονός ότι οι άνθρωποι στις μεγάλες γιορτές είναι ψυχολογικά διατεθειμένοι να καταφύγουν στην τύχη, οπότε σε τέτοιες περιόδους παίζονται μεγαλύτερα ποσά και αλλάζει συμπεριφορά και το παιχνίδι.

Εν κατακλείδι, είναι βέβαιο ότι οι περισσότεροι άνθρωποι επιδιώκουν να κερδίσουν κάποιο μεγάλο ποσό που θα κάνει τη ζωή τους ευκολότερη και θα τους φέρει την ποθητή ευτυχία. Υπάρχουν όμως περιπτώσεις, όπου άνθρωποι που κέρδισαν μεγάλα ποσά εν τέλει καταστράφηκαν, διέλυσαν τις οικογένειές τους, τις σχέσεις τους, τις δουλειές τους και άλλα παρόμοια. Με ποιο τρόπο άραγε, αν είμαστε τυχεροί, θα αποφύγουμε τις κακές συνέπειες; Επιγραμματικά, ένας οδηγός για το τι δεν πρέπει να κάνουμε σε μια τέτοια περίπτωση, είναι:

- Να μη χάσουμε το τυχερό δελτίο.
- Να μην το πούμε από την αρχή σε όλους τους γνωστούς μας.
- Να μην αναλάβουμε μόνοι μας τη διαχείριση αυτού του ποσού. Ίσως συμφέρει να προσλάβουμε κάποιον οικονομικό σύμβουλο.
- Να μην αρχίσουμε να αγοράζουμε τα πάντα σε όλους, αλλά ούτε και στον εαυτό μας.
- Να μην αφήσουμε τα χρέη μας απλήρωτα.
- Και βέβαια, να μην τρελαθούμε!

ABSTRACT

The word Jackpot was previously only known to Casino players. In recent years, however, especially after the introduction of games with numbers, such as LOTTO, JOKER, etc, it has become part of everyday life and it is a matter of great concern to the citizens, who occasionally ask questions like why so many consecutive jackpots occur, why Jackpots are mostly happening in big celebrations, whether they are honest or not, and so on.

In this paper, I tried to give an explanation for such questions using basic principles of Statistics and taking data from the JOKER website, which is accessible to everyone concerned.

ΑΝΑΦΟΡΕΣ

Η επίσημη σελίδα του Τζόκερ: http://www.opap.gr/el/web/guest/joker-draw-results

Διάφορες άλλες ιστοσελίδες σχετικές με το Τζόκερ.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 30ου Πανελληνίου Συνεδρίου Στατιστικής (2017), σελ.57-69

ΕΚΤΙΜΗΣΗ ΤΟΥ ΣΥΝΤΕΛΕΣΤΗ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ ΑΠΟ ΔΕΔΟΜΕΝΑ ΔΙΑΚΡΙΤΗΣ ΟΜΟΙΟΜΟΡΦΗΣ ΚΑΤΑΝΟΜΗΣ

Παπατσούμα Ιωάννα, Φαρμάκης Νικόλαος
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
ioannapapatsouma@gmail.com, farmakis@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία αξιολογείται η αξιοπιστία του εκτιμητή του συντελεστή μεταβλητότητας (ΣΜ), μέσα από τυχαίο δείγμα στοιχείων του πληθυσμού *Π*, όταν η αντίστοιχη τυχαία μεταβλητή ακολουθεί στον πληθυσμό διακριτή ομοιόμορφη κατανομή *DU{0,1, ..., N-1}*. Εξετάζονται περιπτώσεις δειγματοληψίας με επανάθεση και χωρίς επανάθεση και υπολογίζεται το ποσοστό των τιμών του εκτιμητή που βρίσκεται εντός των ορίων μεταβολής του αντίστοιχου θεωρητικού συντελεστή μεταβλητότητας, του πληθυσμού ($\frac{\sqrt{3}}{3} < \Sigma M \leq 1$). Η αξιοπιστία του εκτιμητή αποδίδεται με το ποσοστό των δειγμάτων που δίνουν τιμή του συντελεστή μεταβλητότητας εντός των ορίων του πληθυσμιακού συντελεστή μεταβλητότητας. Με τη μελέτη των ανωτέρω δειγμάτων διαπιστώνεται ότι η αξιοπιστία του δειγματικού συντελεστή μεταβλητότητας αυξάνεται παράλληλα με το μέγεθος του δείγματος. Η συνολική εικόνα που δημιουργείται δίνει μια καλή ιδέα για το «κατά πόσο είναι γενικά αξιόπιστος ο δειγματικός συντελεστής μεταβλητότητας».

Λέξεις Κλειδιά: Συντελεστής μεταβλητότητας, διακριτή ομοιόμορφη κατανομή, δείγμα, αξιοπιστία

AMS ταξινόμηση: 62D05, 62E17

1. ΕΙΣΑΓΩΓΗ

Ο συντελεστής μεταβλητότητας (ΣΜ) είναι ένα από τα σημαντικότερα και χρησιμότερα μέτρα διασποράς που συναντάμε στη βιβλιογραφία (Farmakis, 2003). Έχει χρησιμοποιηθεί ευρέως σε διάφορους επιστημονικούς τομείς, όπως στον τομέα της Ιατρικής Επιστήμης (Usui et al., 2017), της Αναλογιστικής Επιστήμης (Broverman, 2001; Dickson et al., 2013) και σε πολλούς άλλους. Ανάλογα με τις

ανάγκες κάθε έρευνας συναντάται είτε υψωμένος στο τετράγωνο είτε με τη μορφή του αντιστρόφου του (Chaturvedi & Rani, 1996).

Η βασική «καλή» ιδιότητα του συντελεστή μεταβλητότητας είναι ότι είναι αδιάστατο μέγεθος, δηλαδή μια αριθμητική τιμή ανεξάρτητη από τις μονάδες μέτρησης των δεδομένων (*unit free*). Επιτρέπει τη σύγκριση της μεταβλητότητας δύο ή περισσοτέρων συνόλων μετρήσεων που έχουν διαφορετικές μονάδες μέτρησης. Με αυτόν τον τρόπο δίνεται στον ερευνητή μια πανοραμική εικόνα δύο ή περισσοτέρων μεταβλητών ταυτοχρόνως. Επιτρέπει τη σύγκριση της μεταβλητότητας δύο ή περισσοτέρων συνόλων μετρήσεων που έχουν την ίδια μονάδα μέτρησης, αλλά αρκετά διαφορετικούς μέσους όρους ή/και τυπικές αποκλίσεις. Εκφράζει την ομοιογένεια ενός πληθυσμού, καθώς και την ακρίβεια ενός Πειραματικού Σχεδίου. Ακόμη, είναι γνωστό ότι ο δειγματικός συντελεστής μεταβλητότητας δεν είναι αμερόληπτος εκτιμητής του πληθυσμιακού συντελεστή μεταβλητότητας. Οι (Breunig, 2001), (Reed et al., 2002), (Mahmoudvand & Hassani, 2009) και (Sokal & Rohlf, 2012) προτείνουν τη χρήση αμερόληπτων δεικτών.

Με αφορμή την τελευταία ιδιότητα του συντελεστή μεταβλητότητας, αλλά και τη χρησιμότητά του, αξιολογείται στην παρούσα εργασία η αξιοπιστία του μέσω της σχέσης που συνδέει τον δειγματικό με τον πληθυσμιακό συντελεστή. Πιο συγκεκριμένα, αξιολογείται η αξιοπιστία του μέσα από τα δεδομένα μιας διακριτής τ.μ. Το ενδιαφέρον της μελέτης των διακριτών τ.μ. έγκειται στο ότι η μελέτη είναι θεωρητική και μπορεί να γίνει εξαντλητική, επειδή έχουμε μικρό όγκο δεδομένων. Ένα ακόμα πλεονέκτημα της διακριτής τ.μ. είναι ότι μία συνεχής μεταβλητή μπορεί να μετατραπεί σε διακριτή με κατάλληλη διαμέριση του πεδίου τιμών της. Μια προέκταση, επίσης, της μελέτης για μεγαλύτερο πλήθος τιμών μπορεί να δώσει πρόβλεψη και για την οριακή έστω συμπεριφορά της συνεχούς τ.μ., όταν το πλήθος των τιμών μεγαλώνει πολύ ή/και απειρίζεται.

Θα μελετήσουμε, λοιπόν, την περίπτωση της διακριτής τ.μ. X, η οποία ακολουθεί διακριτή ομοιόμορφη κατανομή (*discrete uniform distribution*) $DU\{0,1,\dots,N-1\}$. Έστω X μία τ.μ. η οποία παίρνει τις τιμές $\alpha, \alpha+\omega, \alpha+2\omega, \dots, \alpha+(N-1)\omega$, με την ίδια πιθανότητα. Η πιθανότητα να εμφανιστεί τυχαία μία από τις τιμές της τ.μ. X δίνεται από τη σχέση:

$$f_x(x) = \frac{1}{N}, \qquad x = \alpha, \alpha+\omega \dots, \alpha+(N-1)\cdot\omega$$

Η μέση τιμή και η διασπορά της τ.μ. X (Κολυβά-Μαχαίρα & Μπόρα-Σέντα, 2013 και αλλού) δίνονται, αντίστοιχα, από τις σχέσεις:

$$EX = \alpha + (N-1)\cdot\left(\frac{\omega}{2}\right)$$

$$VarX = \left(\frac{\omega^2}{12}\right)\cdot N\cdot(N+1)$$

Ο συντελεστής μεταβλητότητας της τ.μ. X, με την υπόθεση ότι $\alpha = 0$, δίνεται από τη σχέση:

$$\Sigma M = \sqrt{N \cdot \frac{(N+1)}{(N-1)^2}} \cdot \frac{\sqrt{3}}{3}$$

Παρατηρούμε ότι ο συντελεστής μεταβλητότητας είναι ανεξάρτητος του βήματος ω και ότι, καθώς το μέγεθος του πληθυσμού αυξάνεται, η τιμή του συντελεστή μεταβλητότητας τείνει οριακά στην τιμή $\frac{\sqrt{3}}{3} = 0{,}5774$. Παραγωγίζουμε την παραπάνω σχέση και προκύπτει:

$$\left(\frac{\sqrt{3}}{3}\right) \cdot \frac{-3N-1}{2\left((N-1)^2 \cdot \sqrt{N^2+N}\right)} < 0, \forall \, \mathbf{N} > \mathbf{1}$$

συνεπώς, ο συντελεστής μεταβλητότητας είναι μία γνησίως φθίνουσα συνάρτηση. Είναι, όμως, και άνω φραγμένη και συγκεκριμένα $CV \leq 1, \forall\, N > 2$.

Στην παρούσα εργασία, η αξιοπιστία του εκτιμητή του συντελεστή μεταβλητότητας αποδίδεται τελικά με το ποσοστό των δειγμάτων που δίνουν τιμή του συντελεστή μεταβλητότητας εντός των ορίων του πληθυσμιακού συντελεστή μεταβλητότητας $\frac{\sqrt{3}}{3} < \boldsymbol{\Sigma M \leq 1}$ (Mahmoudvand et al., 2007) ή προσεγγιστικά στο διάστημα $\mathbf{(0,5774, 1]}$.

2. ΜΕΘΟΔΟΣ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Η μέθοδος δειγματοληψίας που επιλέχθηκε είναι η **απλή τυχαία δειγματοληψία** (ΑΤΔ) (Φαρμάκης, 2016), δηλαδή κάθε στοιχείο του πληθυσμού έχει την ίδια πιθανότητα να συμπεριληφθεί στο δείγμα με οποιοδήποτε άλλο στοιχείο του πληθυσμού. Κατά τη διαδικασία της ΑΤΔ θεωρείται συνήθως ότι το εκάστοτε επιλεγόμενο στοιχείο δεν επιστρέφει στον πληθυσμό ώστε να μπορεί να επιλεγεί ξανά. Αυτή η δειγματοληψία είναι γνωστή ως δειγματοληψία χωρίς επανάθεση. Υπάρχει, όμως, περίπτωση το κάθε στοιχείο μόλις επιλεγεί και καταγραφεί η τιμή του, στην προκειμένη περίπτωση η τιμή της τ.μ. που μελετούμε, να επιστρέφει και πάλι στον πληθυσμό ώστε να μπορεί να επανεκλεγεί. Η διαδικασία αυτή ονομάζεται δειγματοληψία με επανάθεση.

Έχει αποδειχθεί ότι $Var\bar{x}_{\varepsilon} > Var\bar{x}$, όπου $\bar{x}_{\varepsilon}$ η μέση του δείγματος με επανάθεση και $\bar{x}$ η μέση τιμή του δείγματος χωρίς επανάθεση και συγκεκριμένα είναι κατά $\frac{N-1}{N-n}$ μεγαλύτερη (Φαρμάκης, 2016), όπου *N* και *n* τα μεγέθη του πληθυσμού και του δείγματος, αντίστοιχα. Επομένως, στην περίπτωση της τυχαίας δειγματοληψία με επανάθεση έχουμε λιγότερη πληροφορία από δείγμα μεγέθους n.

Το πλήθος των διαφορετικών δειγμάτων που λαμβάνονται κατά τη δειγματοληψία χωρίς επανάθεση είναι ίσο με το πλήθος των συνδυασμών N ανά n, το οποίο συμβολίζεται με $\binom{N}{n}$, ενώ κατά τη δειγματοληψία με επανάθεση είναι ίσο με N^n.

Έστω $X = (x_1, x_2, \dots, x_n)$ ένα οποιοδήποτε πεπερασμένο σύνολο και ας υποθέσουμε ότι θέλουμε να χρησιμοποιήσουμε το στοιχείο x_1 συνολικά k_1 φορές, το στοιχείο x_2 συνολικά k_2 φορές και ούτω καθεξής, όπου $k_1 + k_2 + k_3 + \dots = n$. Το πλήθος των επαναληπτικών μεταθέσεων είναι ίσο με $n!/(k_1!k_2!k_3!\dots)$, χωρίς να αποκλείεται κάποια από τα $k_i, i = 1, 2, \dots, n$, να είναι ίσα με 1. Για παράδειγμα, αν έχουμε τρεις φορές "1", μία φορά "2" και δύο φορές "0", τότε έχουμε $6!/(3!1!2!) = 60$ 6-ψήφιους αριθμούς γραμμένους με αυτά τα ψηφία, δηλαδή 60 επαναληπτικές μεταθέσεις.

3. ΕΛΕΓΧΟΣ ΑΞΙΟΠΙΣΤΙΑΣ

Στον έλεγχο αξιοπιστίας του δειγματικού εκτιμητή του συντελεστή μεταβλητότητας ακολουθούνται τα εξής βήματα:

i. καταγραφή των δειγμάτων ως n -διάστατων διανυσμάτων $(x_1, x_2, \dots, x_n)$
ii. υπολογισμός του πλήθους των δειγμάτων
iii. υπολογισμός της μέσης τιμής κάθε δείγματος
iv. υπολογισμός της διασποράς κάθε δείγματος
v. υπολογισμός της τυπικής απόκλισης κάθε δείγματος
vi. υπολογισμός του συντελεστή μεταβλητότητας κάθε δείγματος
vii. υπολογισμός του ποσοστού (επί τοις εκατό) του αριθμού των δειγμάτων όπου ο συντελεστής μεταβλητότητας ανήκει στο διάστημα $\left(\frac{\sqrt{3}}{3}, 1\right]$.
Το ποσοστό αυτό στο εξής θα ονομάζεται χάριν συντομίας «ποσοστό απόδοσης».

3.1 $DU\{0, 1, 2\}$

Αρχικά εξετάστηκε η αξιοπιστία του δειγματικού συντελεστή μεταβλητότητας για την τ.μ. *X* που ακολουθεί διακριτή ομοιόμορφη κατανομή $DU\{0, 1, 2\}$, δηλαδή για $N = 3$.

Για λόγους οικονομίας χώρου, περιγράφεται εκτενώς η περίπτωση των δειγμάτων μεγέθους $n = 6$, τα οποία λαμβάνονται κατά τη δειγματοληψία με επανάθεση. Το καθένα από αυτά τα δείγματα παριστάνεται στον Πίνακα 1 με τη μορφή 6-διάστατου διανύσματος $(x_1, x_2, \dots, x_6)$, οι συντεταγμένες του οποίου μπορούν να πάρουν τις τιμές 0, 1 ή/και 2 καθ' όλους τους δυνατούς τρόπους. Με n_i συμβολίζεται ο αριθμός των επαναληπτικών μεταθέσεων που μπορούν να σχηματιστούν με τις τιμές που δίνονται σε κάθε σειρά του πίνακα.

Πίνακας 1. *Περιγραφή των δειγμάτων και τιμές του ΣΜ για N=3 & n=6*

x_1	x_2	x_3	x_4	x_5	x_6	n_i	ΣΜ

0	**0**	**0**	**0**	**0**	**0**	**1**	**1,0000**
1	0	0	0	0	0	6	2,4495
2	0	0	0	0	0	6	2,4495
1	1	0	0	0	0	15	1,5492
1	2	0	0	0	0	30	1,6733
2	2	0	0	0	0	15	1,5492
1	1	1	0	0	0	20	1,0954
1	1	2	0	0	0	60	1,2247
1	2	2	0	0	0	60	1,1798
2	2	2	0	0	0	20	1,0954
1	**1**	**1**	**1**	**0**	**0**	**15**	**0,7746**
1	**1**	**1**	**2**	**0**	**0**	**60**	**0,9033**
1	**1**	**2**	**2**	**0**	**0**	**90**	**0,8944**
1	**2**	**2**	**2**	**0**	**0**	**60**	**0,8427**
2	**2**	**2**	**2**	**0**	**0**	**15**	**0,7746**
1	1	1	1	1	0	6	0,4899
1	**1**	**1**	**1**	**2**	**0**	**30**	**0,6325**
1	**1**	**1**	**2**	**2**	**0**	**60**	**0,6452**
1	**1**	**2**	**2**	**2**	**0**	**60**	**0,6124**
1	2	2	2	2	0	30	0,5578
2	2	2	2	2	0	6	0,4899
1	1	1	1	1	1	1	0,0000
1	1	1	1	1	2	6	0,3499
1	1	1	1	2	2	15	0,3873
1	1	1	2	2	2	20	0,3651
1	1	2	2	2	2	15	0,3098
1	2	2	2	2	2	6	0,2227
2	2	2	2	2	2	1	0,0000
ΣΥΝΟΛΟ						**729**	

Συνολικά, λαμβάνουμε $N^n = 3^6 = 729$ δείγματα μεγέθους 6. Παρατηρούμε ότι σε 391 δείγματα, τα οποία επισημαίνονται με έντονη γραφή (bold), ο συντελεστής μεταβλητότητας ανήκει στο διάστημα $\left(\frac{\sqrt{3}}{3}, 1\right]$.

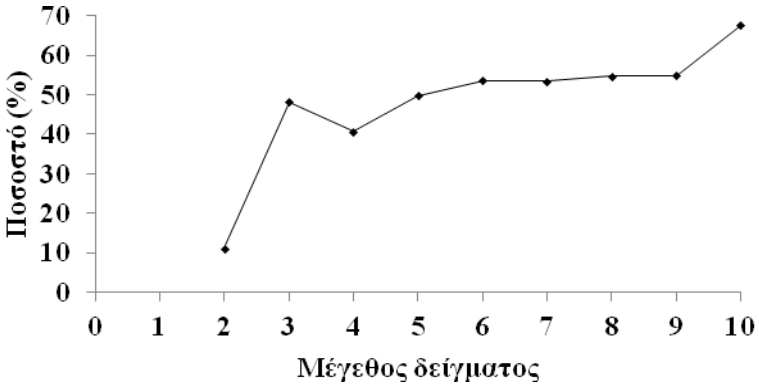
Επαναλαμβάνουμε την ίδια διαδικασία για κάθε δείγμα με επανάθεση μεγέθους $n = 2, 3, 4, 5, 7, 8, 9$ και 10 και καταγράφουμε τον αριθμό των συνολικών δειγμάτων, καθώς και τον αριθμό και το ποσοστό των δειγμάτων όπου ο συντελεστής μεταβλητότητας ανήκει στο διάστημα $\left(\frac{\sqrt{3}}{3}, 1\right]$ (Πίνακας 2).

Πίνακας 2. *Αποτελέσματα δειγματοληψίας με επανάθεση για N=3 & n=2, 3, …, 10*

Μέγεθος δείγματος	**Αριθμός δειγμάτων**	**Αριθμός δειγμάτων όπου $\frac{\sqrt{3}}{3} < \Sigma M \leq 1$**	**Ποσοστό απόδοσης (%)**
2	9	1	**11,11**
3	27	13	**48,15**
4	81	33	**40,74**
5	243	121	**49,79**
6	729	391	**53,64**
7	2187	1170	**53,50**
8	6561	3585	**54,64**
9	19683	10795	**54,84**
10	59049	39871	**67,52**

Παρατηρείται μία τάση το ποσοστό των δειγματικών τιμών του συντελεστή μεταβλητότητας που είναι μέσα στα όρια του πληθυσμιακού συντελεστή μεταβλητότητας να αυξάνεται παράλληλα με το μέγεθος του δείγματος (Σχήμα 1).

Σχήμα 1. *Ποσοστό απόδοσης (%) για N=3 & n=2, 3, …, 10*



3.2 $DU\{0,1,2,3\}$

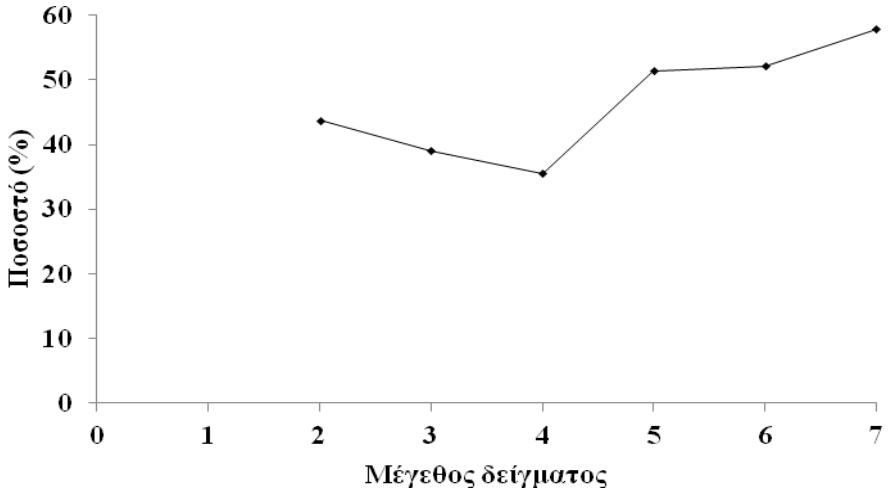
Στη συνέχεια εξετάστηκε η αξιοπιστία του δειγματικού συντελεστή μεταβλητότητας για την τ.μ. X που ακολουθεί διακριτή ομοιόμορφη κατανομή $DU\{0,1,2,3\}$, δηλαδή για $N = 4$. Στον Πίνακα 3 παρουσιάζονται τα αποτελέσματα της δειγματοληψίας με επανάθεση για $n = 2,3,4,5,6$ και 7.

Πίνακας 3. *Αποτελέσματα δειγματοληψίας με επανάθεση για N=4 & n=2, 3, …, 7*

Μέγεθος δείγματος	**Αριθμός δειγμάτων**	**Αριθμός δειγμάτων όπου $\frac{\sqrt{3}}{3} < \Sigma M \leq 1$**	**Ποσοστό απόδοσης (%)**
2	16	7	**43,75**
3	64	25	**39,00**
4	256	91	**35,55**
5	1024	526	**51,37**
6	4096	2137	**52,17**
7	16384	9479	**57,86**

Είναι φανερή και εδώ η τάση να αυξάνεται το ποσοστό των δειγματικών τιμών του συντελεστή μεταβλητότητας που είναι μέσα στα όρια του πληθυσμιακού συντελεστή μεταβλητότητας (Σχήμα 2).

Σχήμα 2. *Ποσοστό απόδοσης (%) για N=4 & n=2, 3, …, 7*



3.3 $DU\{0, 1, 2, 3, 4\}$

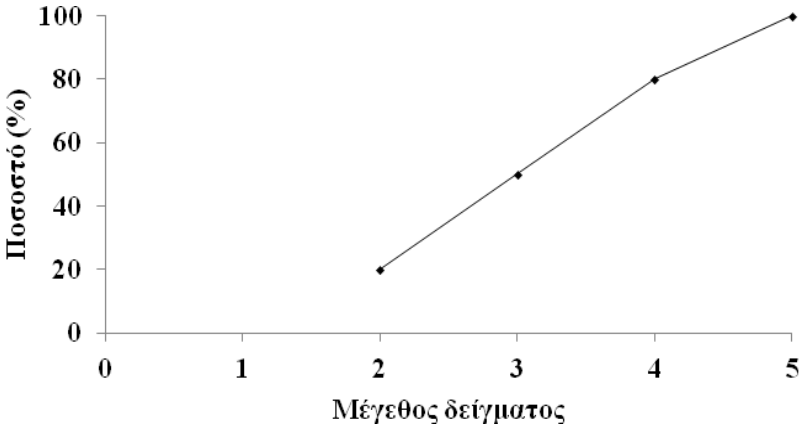
Στη συνέχεια εξετάστηκε η αξιοπιστία του δειγματικού συντελεστή μεταβλητότητας για την τ.μ. X που ακολουθεί διακριτή ομοιόμορφη κατανομή $DU\{0, 1, 2, 3, 4\}$, δηλαδή για $N = 5$. Στον Πίνακα 4 παρουσιάζονται τα αποτελέσματα της δειγματοληψίας χωρίς επανάθεση για $n = 2,\ 3,\ 4$ και 5.

Πίνακας 4. *Αποτελέσματα δειγματοληψίας χωρίς επανάθεση για N=5 & n=2, 3, 4, 5*

Μέγεθος δείγματος	**Αριθμός δειγμάτων**	**Αριθμός δειγμάτων όπου $\frac{\sqrt{3}}{3} < \Sigma M \leq 1$**	**Ποσοστό απόδοσης (%)**
2	10	2	**20**
3	10	5	**50**
4	5	4	**80**
5	1	1	**100**

Παρατηρούμε ότι υπάρχει μόνο ένα δείγμα μεγέθους $n = 5$, ο ίδιος ο πληθυσμός. Στην περίπτωση αυτή, οι τιμές του δειγματικού και του πληθυσμιακού συντελεστή μεταβλητότητας συμπίπτουν και είναι ίσες με $0{,}7906$.

Σχήμα 3. *Ποσοστό απόδοσης (%) για N=5 & n=2, 3, 4, 5*



Το ποσοστό των δειγμάτων, όπου ο συντελεστής μεταβλητότητας ανήκει στο διάστημα $\left(\frac{\sqrt{3}}{3}, 1\right]$, έχει και εδώ αύξουσα και έντονα γραμμική τάση σε σχέση με το μέγεθος του δείγματος (Σχήμα 3).

3.4 $DU\{0, 1, 2, 3, 4, 5, 6, 7\}$

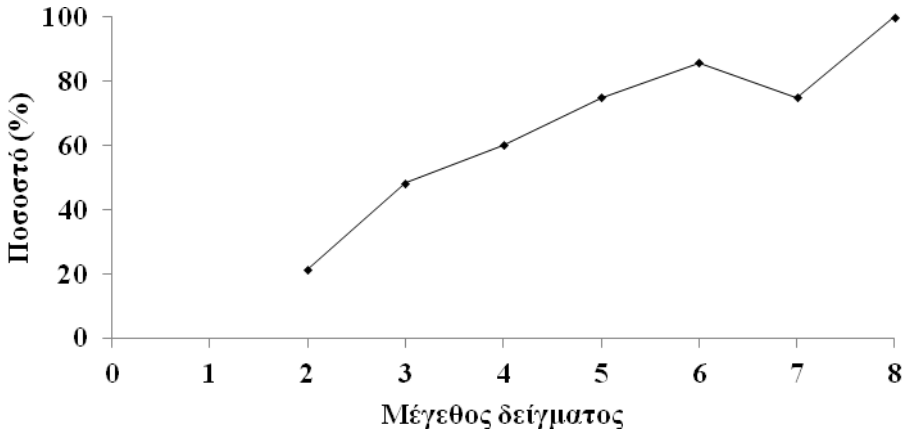
Τέλος, εξετάστηκε η αξιοπιστία του δειγματικού συντελεστή μεταβλητότητας για την τ.μ. X που ακολουθεί διακριτή ομοιόμορφη κατανομή $DU\{0, 1, 2, 3, 4, 5, 6, 7\}$, δηλαδή για $N = 8$. Στον Πίνακα 5 παρουσιάζονται τα αποτελέσματα της δειγματοληψίας χωρίς επανάθεση για $n = 2,\ 3,\ 4,\ 5,\ 6,\ 7$ και 8.

Πίνακας 5. *Αποτελέσματα δειγματοληψίας χωρίς επανάθεση για N=8 & n=2, 3, …, 8*

Μέγεθος δείγματος	**Αριθμός δειγμάτων**	**Αριθμός δειγμάτων όπου $\frac{\sqrt{3}}{3} < \Sigma M \leq 1$**	**Ποσοστό απόδοσης (%)**
2	28	6	**21,43**
3	56	27	**48,21**
4	70	42	**60,00**
5	56	42	**75,00**
6	28	24	**85,71**
7	8	6	**75,00**
8	1	1	**100,00**

Παρατηρούμε ότι υπάρχει μόνο ένα δείγμα μεγέθους $n = 8$, ο ίδιος ο πληθυσμός. Στην περίπτωση αυτή, οι τιμές του δειγματικού και του πληθυσμιακού συντελεστή μεταβλητότητας συμπίπτουν και είναι ίσες με $0{,}6999$.

Σχήμα 4. *Ποσοστό απόδοσης (%) για N=8 & n=2, 3, …, 8*



Γενικά, το ποσοστό απόδοσης έχει και εδώ μια αύξουσα τάση σε σχέση με το μέγεθος του δείγματος (Σχήμα 4).

4. ΣΥΣΧΕΤΙΣΗ ΜΕΓΕΘΟΥΣ ΤΟΥ ΔΕΙΓΜΑΤΟΣ ΜΕ ΠΟΣΟΣΤΟ ΑΠΟΔΟΣΗΣ

Μετά την ολοκλήρωση του ελέγχου αξιοπιστίας, εξετάστηκε ο ρυθμός αύξησης του ποσοστού απόδοσης για αύξηση του μεγέθους, *n*, του δείγματος κατά μία μονάδα.

Για λόγους οικονομίας χώρου, περιγράφεται εκτενώς η περίπτωση για $N = 3$ και δειγματοληψία με επανάθεση, όπου η τιμή του συντελεστή συσχέτισης ξεπερνάει το $0{,}8$ (Πίνακας 6), δηλαδή με άλλα λόγια τουλάχιστον το 64% των μεταβολών του ποσοστού απόδοσης οφείλεται στη μεταβολή του μεγέθους του δείγματος. Η τιμή του συντελεστή συσχέτισης υποδεικνύει ότι υπάρχει γραμμική σχέση μεταξύ του ποσοστού απόδοσης και του μεγέθους του δείγματος και η ευθεία της γραμμικής παλινδρόμησης δίνεται από τη σχέση:

$$\hat{y} = b_0 + b_1 \cdot n$$

όπου $\hat{y}$ η εκτίμηση του ποσοστού απόδοσης, n το μέγεθος του δείγματος και b_0, b_1 οι συντελεστές παλινδρόμησης. Η τιμή του συντελεστή b_1 προκύπτει από τα ζεύγη τιμών μέγεθος δείγματος – ποσοστό απόδοσης (%) του Πίνακα 2 και είναι κατά μέσο όρο λίγο μεγαλύτερη από το 2. Αυτό μας επιτρέπει να υποθέσουμε μια αύξηση του

ποσοστού απόδοσης κατά 2% περίπου σε κάθε αύξηση του μεγέθους n του δείγματος κατά μία μονάδα. Άρα, με μια αρχική αξιοπιστία του συντελεστή μεταβλητότητας γύρω στο 35% χρειαζόμαστε μέγεθος δείγματος 35 με 40 στοιχεία για μεγάλη αξιοπιστία που να πλησιάζει το 100%. Ο συντελεστής b_1 που προκύπτει από παλινδρόμηση με στοιχεία που περιλαμβάνουν και μεγέθη δείγματος $n > 10$ βαίνει οριακά μειούμενος. Συνεπώς και η μέση τιμή θα πέσει και κάτω από 2.

Στον Πίνακα 6 παρουσιάζονται οι τιμές των συντελεστών συσχέτισης και προσδιορισμού για όλες τις περιπτώσεις δειγματοληψίας που εξετάστηκαν. Όλες οι τιμές των συντελεστών συσχέτισης είναι θετικές, επομένως υπάρχει θετική γραμμική εξάρτηση μεταξύ των δύο μεταβλητών (Κολυβά-Μαχαίρα & Μπόρα-Σέντα, 2013), του ποσοστού απόδοσης και του μεγέθους του δείγματος, είτε αυτό λαμβάνεται με επανάθεση είτε χωρίς επανάθεση.

Πίνακας 6. *Τιμές των συντελεστών συσχέτισης και προσδιορισμού*

	Με επανάθεση		**Χωρίς επανάθεση**	
Μέγεθος πληθυσμού	3	4	5	8
Συντελεστής συσχέτισης	0,8490	0,7837	0,9959	0,9361
Συντελεστής προσδιορισμού	0,7208	0,6142	0,9918	0,8763

Η τιμή του συντελεστή συσχέτισης για $N = 5$ και δειγματοληψία χωρίς επανάθεση ($r = 0{,}9959$) επιβεβαιώνει ότι υπάρχει «σχεδόν» τέλεια γραμμική σχέση μεταξύ του μεγέθους του δείγματος και του ποσοστού απόδοσης, όπως είδαμε και στο Σχήμα 3. Η αντίστοιχη τιμή του συντελεστή προσδιορισμού ($r^2 = 0{,}9918$) επιβεβαιώνει την πολύ καλή προσαρμογή της ευθείας παλινδρόμησης στα δεδομένα της διακριτής ομοιόμορφης κατανομής.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα δείγματα που λαμβάνονται με επανάθεση μας δίνουν λιγότερη πληροφορία από αυτήν που δίνουν τα δείγματα ίδιου μεγέθους που λαμβάνονται χωρίς επανάθεση.

Καθώς αυξάνεται ο αριθμός *N* των στοιχείων του Πληθυσμού, παρατηρείται αύξηση του ποσοστού των δειγμάτων που λαμβάνονται τόσο με επανάθεση, όσο και χωρίς επανάθεση και δίνουν τιμή του δειγματικού συντελεστή μεταβλητότητας εντός των ορίων του πληθυσμιακού συντελεστή μεταβλητότητας.

Η τιμή του συντελεστή συσχέτισης μας δείχνει ότι υπάρχει εντονότερη θετική γραμμική συσχέτιση μεταξύ του μεγέθους n των δειγμάτων χωρίς επανάθεση και του ποσοστού των δειγμάτων, όπου ο συντελεστής μεταβλητότητας παίρνει τιμή εντός των ορίων του πληθυσμιακού συντελεστή μεταβλητότητας, συγκριτικά με τα δείγματα που λαμβάνονται με επανάθεση.

Μοντέλα γραμμικής παλινδρόμησης προβλέπουν ότι δείγματα μεγέθους $n > 40$ δίνουν υψηλά ποσοστά δειγμάτων με τιμή του συντελεστή μεταβλητότητας εντός του διαστήματος $\left(\frac{\sqrt{3}}{3}, \mathbf{1}\right]$.

ABSTRACT

In the present paper we evaluate the reliability of the estimated coefficient of variation (CV) by a random sample from the population, when the random variable follows the discrete uniform distribution *DU{0,1, ..., N-1}*. Samples both with replacement and without replacement are examined and the percentage of the values of the estimator that lie within the corresponding population coefficient of variation's bounds $\left(\frac{\sqrt{3}}{3} < CV \leq \mathbf{1}\right)$ is calculated. The reliability of the estimator is attributed to the percentage of the samples where the coefficient of variation lies within the population coefficient of variation's bounds. The study of the above samples shows that the reliability of the sampling coefficient of variation increases in parallel with the sample size. The overall study gives a good idea of "whether the sampling coefficient of variation is generally reliable".

Keywords: Coefficient of variation, discrete uniform distribution, sample, reliability

AMS Classification: 62D05, 62E17

ΑΝΑΦΟΡΕΣ

Κολυβά-Μαχαίρα Φ., Μπόρα-Σέντα Ε. (2013). *Στατιστική: Θεωρία και Εφαρμογές*. Β΄ Έκδοση. Εκδόσεις Ζήτη, Θεσσαλονίκη.

Φαρμάκης Ν. (2001). *ΣΤΑΤΙΣΤΙΚΗ, Περιληπτική Θεωρία-Ασκήσεις*. Β΄ Έκδοση. Εκδόσεις Α&Π Χριστοδουλίδη, Θεσσαλονίκη.

Φαρμάκης Ν. (2016). *Εισαγωγή στη Δειγματοληψία.* Αφοί Κυριακίδη Εκδόσεις Α.Ε., Θεσσαλονίκη.

Breunig, R. (2001). An almost unbiased estimator of the coefficient of variation. *Economics Letters* **70**:15–19.

Broverman S. A. (2001). Actex study manual, Course 1, Examination of the Society of Actuaries, Exam 1 of the Casualty Actuarial Society, 2001 ed. Winsted, CT: Actex Publications.

Chaturvedi A., Rani U. (1996). Fixed-Width Confidence Interval Estimation of the Inverse Coefficient of Variation in a Normal Distribution. *Microelectron. Relab.*, **36**(9):1305-1308.

Dickson D. C. M., Hardy M. R., Waters H. R. (2013). *Actuarial Mathematics for Life Contingent Risks*, 2nd edition. Cambridge University Press.

Farmakis N. (2003). Estimation of Coefficient of Variation: Scaling of Symmetric Continuous Distributions, *Statistics in Transition*, **6**(1):83-96.

Mahmoudvand R., Hassani H., Wilson R. (2007). Is the Sample Coefficient of Variation a Good Estimator for the Population Coefficient of Variation?, *World Applied Sciences Journal*, **2**(5):519-522.

Mahmoudvand, R. Hassani, H. (2009). Two new confidence intervals for the coefficient of variation in a normal distribution. *Journal of Applied Statistics* **36**:429–442.

Sokal, R.R. Rohlf, F.J. (2012). *Biometry: the principles and practice of statistics in biological research*, 4th ed. W. H. Freeman and Co.: New York.

Usui K., Mori H., Tachi T., Matsumura T., Mori K., Takeda A., Noguchi Y., Yoshimura T., Teramachi H. (2017). A rapid method to screen poisoning causative agents in an acute care hospital in Japan. *Journal of Clinical Pharmacy and Therapeutics*, **42**(4):454–460.

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά του 30ου Πανελλήνιου Συνέδριου Στατιστικής (2017), σελ. 70−79

ΑΚΡΙΒΕΙΣ ΕΛΕΓΧΟΙ ΓΙΑ ΤΟΝ ΛΟΓΟ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΚΛΙΜΑΚΑΣ ΔΥΟ ΚΑΤΑΝΟΜΩΝ LAPLACE

Μ. Ταφιάδη, Γ. Ηλιόπουλος
Πανεπιστήμιο Πειραιώς
{mtafiadi, geh}@unipi.gr

ΠΕΡΙΛΗΨΗ

Θεωρούμε ακριβείς ελέγχους για να εξετάσουμε την ισότητα του λόγου παραμέτρων κλίμακας, δύο πληθυσμών Laplace βασιζόμενοι σε αντίστοιχα ανεξάρτητα τυχαία δείγματα. Οι στατιστικές συναρτήσεις βασίζονται είτε στους εκτιμητές μέγιστης πιθανοφάνειας, είτε στους βέλτιστους γραμμικούς αμερόληπτους εκτιμητές. Δεσμεύοντας σε συγκεκριμένες ποσότητες, εκφράζουμε τις ακριβείς κατανομές τους ως μείξεις κατανομών λόγων γραμμικών συνδυασμών ανεξάρτητων τυπικών εκθετικών τυχαίων μεταβλητών. Αυτό μας επιτρέπει να βρούμε σε κλειστή μορφή την συνάρτηση κατανομής των στατιστικών συναρτήσεων και να υπολογίσουμε ποσοστιαία σημεία. Κατασκευάζουμε αμερόληπτους ελέγχους και αμερόληπτα διαστήματα εμπιστοσύνης για τον λόγο των δύο παραμέτρων κλίμακας τα οποία βασίζονται στους παραπάνω εκτιμητές. Οι ακριβείς διαδικασίες παρουσιάζονται μέσω ενός παραδείγματος με πραγματικά δεδομένα.

Λέξεις κλειδιά: Κατανομή Laplace, λόγος παραμέτρων κλίμακας, ακριβείς έλεγχοι, εκτιμητές μέγιστης πιθανοφάνειας, βέλτιστοι γραμμικοί αμερόληπτοι εκτιμητές

1. ΕΙΣΑΓΩΓΗ

Ένα σημαντικό πρόβλημα στην Στατιστική είναι η σύγκριση των παραμέτρων κλίμακας δύο πληθυσμών. Για το πρόβλημα αυτό υπάρχουν ακριβείς παραμετρικοί έλεγχοι μόνο για συγκεκριμένες οικογένειες κατανομών. Για να κάνουμε τέτοιες συγκρίσεις έξω από το πλαίσιο αυτών των κατανομών, θα πρέπει να βασιστούμε είτε σε προσεγγιστικές (π.χ. ασυμπτωτικές) λύσεις, είτε σε μη παραμετρικούς ελέγχους.

Σε αυτή την εργασία συζητάμε ακριβείς ελέγχους για τη σύγκριση των παραμέτρων κλίμακας δύο κατανομών Laplace βασιζόμενοι σε αντίστοιχα ανεξάρτητα τυχαία δείγματα. Οι έλεγχοι βασίζονται στους εκτιμητές μέγιστης πιθανοφάνειας (ΕΜΠ) και στους βέλτιστους γραμμικούς αμερόληπτους εκτιμητές (ΒΓΑΕ). Η συνάρτηση πυκνότητας πιθανότητας της κατανομής Laplace με παράμετρο θέσης $\mu \in \mathbb{R}$ και παράμετρο

κλίμακας $\sigma > 0$, που θα τη συμβολίζουμε με $\mathcal{L}(\mu, \sigma)$, δίνεται από τον εξής τύπο:

$$f(x; \mu, \sigma) = \frac{1}{2\sigma} \mathrm{e}^{-|x-\mu|/\sigma}, \quad x \in \mathbb{R}.$$

Έστω

$$X = (X_1, \ldots, X_{n_1}),\ Y = (Y_1, \ldots, Y_{n_2}), \quad n_1, n_2 \geqslant 2, \tag{1}$$

δύο ανεξάρτητα τυχαία δείγματα από τις κατανομές $\mathcal{L}(\mu_1, \sigma_1), \mathcal{L}(\mu_2, \sigma_2)$, αντίστοιχα, και

$$X_{1:n_1} < X_{2:n_1} < \cdots < X_{n_1:n_1}, \quad Y_{1:n_2} < Y_{2:n_2} < \cdots < Y_{n_2:n_2}$$

οι αντίστοιχες διατεταγμένες στατιστικές συναρτήσεις. Θεωρούμε τον έλεγχο της υπόθεσης

$$H_0 : \sigma_1/\sigma_2 = \rho_0 \quad \text{κατά} \quad H_1 : \sigma_1/\sigma_2 \neq \rho_0.$$

Στα επόμενα θα θεωρήσουμε μόνο την περίπτωση $\rho_0 = 1$. Αν $\rho_0 \neq 1$ τότε το πρόβλημα ανάγεται στην πρώτη περίπτωση πολλαπλασιάζοντας τα X_i με ρ_0. Συζητάμε δύο ακριβείς ελέγχους που βασίζονται στους λόγους αντίστοιχων εκτιμητών των σ_1, σ_2 και συγκεκριμένα των ΕΜΠ και των ΒΓΑΕ. Παρεμπιπτόντως, ο πρώτος από τους δύο συμπίπτει με τον έλεγχο γενικευμένου λόγου πιθανοφανειών. Εκφράζοντας την κατανομή των δύο στατιστικών συναρτήσεων μέσω κλειστών τύπων υπολογίζουμε ακριβώς τα κρίσιμα σημεία των δύο ελέγχων. Αυτό το επιτυγχάνουμε χρησιμοποιώντας ένα αποτέλεσμα των Iliopoulos and Balakrishnan (2011) οι οποίοι ανέπτυξαν ακριβή συμπερασματολογία για τις παραμέτρους μίας κατανομής Laplace.

Στην ενότητα 2 θεωρούμε τους ελέγχους που βασίζοται είτε στους ΕΜΠ, είτε στους ΒΓΑΕ και βρίσκουμε τις ακριβείς κατανομές τους. Στην ενότητα 3 περιγράφουμε τρόπο κατασκευής αμερόληπτων ελέγχων και αμερόληπτων διαστημάτων εμπιστοσύνης. Τέλος, στην ενότητα 4 παρουσιάζουμε όλες τις διαδικασίες μέσω ενός αριθμητικού παραδείγματος.

2. ΕΛΕΓΧΟΙ ΓΙΑ ΤΟ ΛΟΓΟ ΠΑΡΑΜΕΤΡΩΝ ΚΛΙΜΑΚΑΣ

Θεωρούμε τα δεδομένα στην (1) και θέτουμε $n = n_1 + n_2$, $m_i = [(n_i + 1)/2]$, $i = 1, 2$. Η συνάρτηση πιθανοφάνειας του $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ είναι

$$L(\mu_1, \mu_2, \sigma_1, \sigma_2 | X, Y) = \frac{1}{2^{n_1+n_2} \sigma_1^{n_1} \sigma_2^{n_2}}\, e^{-\frac{1}{\sigma_1}\sum_{i=1}^{n_1} |X_i - \mu_1| - \frac{1}{\sigma_2}\sum_{i=1}^{n_2} |Y_i - \mu_2|},$$

$$\mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 > 0,$$

και είναι εύκολο να δει κανείς ότι μεγιστοποιείται στο σημείο $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ όπου $\hat{\mu}_1, \hat{\mu}_2$ είναι οποιεσδήποτε διάμεσοι των X, Y, αντίστοιχα, και

$$\hat{\sigma}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}|X_i - \hat{\mu}_1| = \frac{1}{n_1}\Big(\sum_{i=m_1+1}^{n_1} X_{i:n_1} - \sum_{i=1}^{[n_1/2]} X_{i:n_1}\Big),$$

$$\hat{\sigma}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}|Y_i - \hat{\mu}_2| = \frac{1}{n_2}\Big(\sum_{i=m_2+1}^{n_2} Y_{i:n_2} - \sum_{i=1}^{[n_2/2]} Y_{i:n_2}\Big)$$

είναι οι ΕΜΠ των σ_1, σ_2. Υπό τη μηδενική υπόθεση $\sigma_1 = \sigma_2$ η πιθανοφάνεια μεγιστοποιείται στο σημείο $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{\sigma})$ όπου

$$\hat{\sigma} = \frac{1}{n}\Big\{\sum_{i=1}^{n_1}|X_i - \hat{\mu}_1| + \sum_{i=1}^{n_2}|Y_i - \hat{\mu}_2|\Big\} = \frac{n_1\hat{\sigma}_1 + n_2\hat{\sigma}_2}{n}.$$

Κάνοντας τις απαραίτητες πράξεις καταλήγουμε στο ότι ο γενικευμένος λόγος πιθανοφανειών είναι

$$\frac{L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2|X, Y)}{L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{\sigma}|X, Y)} = \frac{\hat{\sigma}^n}{\hat{\sigma}_1^{n_1}\hat{\sigma}_2^{n_2}} \propto \frac{(n_2/n_1 + \hat{\sigma}_1/\hat{\sigma}_2)^n}{(\hat{\sigma}_1/\hat{\sigma}_2)^{n_1}},$$

και παίρνει μεγάλη τιμή αν ο λόγος

$$T_1 := \hat{\sigma}_1/\hat{\sigma}_2$$

πάρει είτε μικρή είτε μεγάλη τιμή. Επομένως, η $H_0 : \sigma_1 = \sigma_2$ απορρίπτεται για μικρές ή μεγάλες τιμές του λόγου των ΕΜΠ των σ_1, σ_2.

Μιμούμενοι τον παραπάνω λογικό κανόνα θεωρούμε και τον έλεγχο που βασίζεται στον αντίστοιχο λόγο των βέλτιστων γραμμικών αμερόληπτων εκτιμητών (ΒΓΑΕ) $\tilde{\sigma}_1, \tilde{\sigma}_2$ των σ_1, σ_2,

$$T_2 = \tilde{\sigma}_1/\tilde{\sigma}_2.$$

Οι εκτιμητές αυτοί έχουν την μορφή $\sum_{i=1}^{n_1}\beta_{in_1}X_{i:n_1}, \sum_{i=1}^{n_2}\beta_{in_2}Y_{i:n_2}$ με τους συντελεστές β_{in_j} να έχουν δοθεί από τον Govindarajulu (1966).

Προκειμένου να βρούμε κρίσιμα σημεία για τους δύο ελέγχους θα χρειαστούμε την ακριβή κατανομή των δύο στατιστικών συναρτήσεων. Επειδή τόσο οι ΒΓΑΕ όσο και οι ΕΜΠ των σ_1, σ_2 είναι γραμμικοί συνδυασμοί των διατεταγμένων στατιστικών συναρτήσεων που προκύπτουν από τα δύο δείγματα, θα διατυπώσουμε ένα γενικότερο αποτέλεσμα το οποίο καλύπτει και τις δύο:

Πρόταση 1. *Έστω δεδομένα όπως στην* (1) *και* $a_1, \ldots, a_{n_1}, b_1, \ldots, b_{n_2}$ *κάποιες σταθερές. Τότε, για οποιοδήποτε* $x \in \mathbb{R}$,

$$P\Big\{\frac{\sum_{i=1}^{n_1} a_i X_{i:n_1}}{\sum_{i=1}^{n_2} b_i Y_{i:n_2}} \leqslant x\Big\} = \frac{1}{2^n}\sum_{d_1=0}^{n_1}\sum_{d_2=0}^{n_2}\binom{n_1}{d_1}\binom{n_2}{d_2}P\Big\{\frac{\sum_{i=1}^{n_1} s_i(d_1)U_i}{\sum_{i=1}^{n_2} t_i(d_2)V_i} \leqslant \frac{\sigma_2}{\sigma_1}\,x\Big\} \tag{2}$$

όπου $U_1, \ldots, U_{n_1}, V_1, \ldots, V_{n_2}$, *είναι ανεξάρτητες τυπικές εκθετικές τυχαίες μεταβλητές και*

$$s_i(d_1) = \begin{cases} -\dfrac{\sum_{j=1}^{d_1-i+1} a_j}{d_1-i+1}, & i \leqslant d_1, \\ \dfrac{\sum_{j=i}^{n_1} a_j}{n_1-i+1}, & i > d_1, \end{cases} \qquad t_i(d_2) = \begin{cases} -\dfrac{\sum_{j=1}^{d_2-i+1} b_j}{d_2-i+1}, & i \leqslant d_2, \\ \dfrac{\sum_{j=i}^{n_2} b_j}{n_2-i+1}, & i > d_2. \end{cases}$$

Απόδειξη. Είναι γνωστό ότι αν $Z \sim \mathcal{L}(\mu, \sigma)$ τότε η δεσμευμένη κατανομή της $\mu - Z$ δοθέντος ότι $Z \leq \mu$ και η δεσμευμένη κατανομή της $Z - \mu$ δοθέντος ότι $Z \geq \mu$ είναι η εκθετική κατανομή με μέση τιμή σ, $\mathcal{E}(\sigma)$ (βλ. Kotz et al., 2001). Έστω $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} L(\mu, \sigma)$ και $D = \#\{Z_i \leq \mu\}$ (το πλήθος των Z_i που είναι μικρότερα ή ίσα του μ). Αφού το μ είναι η διάμεσος της κατανομής, η D ακολουθεί διωνυμική κατανομή $\mathcal{B}(n, 1/2)$. Από τους Iliopoulos and Balakrishnan (2009) γνωρίζουμε ότι δοθέντος $D = d$ τα μπλοκ $(Z_{1:n}, \ldots, Z_{d:n})$ και $(Z_{d+1:n}, \ldots, Z_{n:n})$ είναι ανεξάρτητα και συνεπώς $(\mu - Z_{1:n}, \ldots, \mu - Z_{d:n}) \overset{\text{d}}{=} (L_{d:d}, \ldots, L_{1:d})$ και $(Z_{d+1:n} - \mu, \ldots, Z_{n:n} - \mu) \overset{\text{d}}{=} (R_{1:n-d}, \ldots, R_{n-d:n-d})$, όπου $L_1, \ldots, L_d$ και $R_1, \ldots, R_{n-d}$ είναι ανεξάρτητες τυχαίες μεταβλητές με κατανομή $\mathcal{E}(\sigma)$. Είναι επίσης γνωστό ότι αν $W_1, \ldots, W_k$ είναι ανεξάρτητες τυχαίες μεταβλητές $\mathcal{E}(\sigma)$, τότε οι τυχαίες μεταβλητές $kW_{1:k}, (k-1)(W_{2:k} - W_{1:k}), \ldots, W_{k:k} - W_{k-1:k}$ είναι ανεξάρτητες με κατανομή επίσης $\mathcal{E}(\sigma)$ (βλ. Arnold et al., 2008). Χρησιμοποιώντας όλα τα παραπάνω συμπεραίνουμε ότι δοθέντος $D = d$ ο γραμμικός συνδυασμός $\sum_{i=1}^{n} c_i Z_i$ έχει την ίδια κατανομή με την τυχαία μεταβλητή $-\sum_{j=1}^{d} \frac{\sum_{i=1}^{d-j+1} c_i}{d-j+1} W_j + \sum_{j=d+1}^{n} \frac{\sum_{i=j}^{n} c_i}{n-j+1} W_j$, όπου $W_1, \ldots, W_n \overset{\text{iid}}{\sim} \mathcal{E}(\sigma)$.

Έστω τώρα $D_1 = \#\{X_i \leqslant \mu_1\}, D_2 = \#\{Y_i \leqslant \mu_2\}$ που λόγω της ανεξαρτησίας των X και Y είναι επίσης ανεξάρτητες με διωνυμικές κατανομές $\mathcal{B}(n_1, 1/2), \mathcal{B}(n_2, 1/2)$, αντίστοιχα. Βάσει των παραπάνω, η κατανομή του λόγου $\sum_{i=1}^{n_1} a_i X_i / \sum_{i=1}^{n_2} b_i Y_i$ δοθέντος $D_1 = d_1$ και $D_2 = d_2$ είναι ίδια με του λόγου $\sum_{i=1}^{n_1} s_i(d_1)\sigma_1 U_i / \sum_{i=1}^{n_2} t_i(d_2)\sigma_2 V_i$ με τους συντελεστές $s_i(d_1), t_i(d_2)$ όπως στην εκφώνηση του θεωρήματος. Αθροίζοντας για όλα τα d_1, d_2 παίρνουμε την (2). □

Η (2) μας λέει ότι η κατανομή του λόγου γραμμικών συνδυασμών των διατεταγμένων στατιστικών συναρτήσεων που προκύπτουν από δύο ανεξάρτητα τυχαία δείγματα από κατανομές Laplace είναι μείξη κατανομών λόγων γραμμικών συνδυασμών ανεξάρτητων εκθετικών τυχαίων μεταβλητών. Οι ΕΜΠ και οι ΒΓΑΕ των σ_1, σ_2 είναι γραμμικές συναρτήσεις των αντίστοιχων διατεταγμένων στατιστικών συναρτήσεων, επομένως η συνάρτηση κατανομής των στατιστικών συναρτήσεων T_1, T_2 έχει την μορφή (2). Επειδή δε οι T_1, T_2 είναι θετικές με πιθανότητα ένα, οι συντελεστές $s_i(d_1), t_i(d_2)$ είναι και αυτοί θετικοί. (Σε διαφορετική περίπτωση θα υπήρχε τουλάχιστον ένα ζεύγος d_1, d_2 έτσι ώστε ο λόγος $\sum s_i(d_1)U_i / \sum t_i(d_2)V_i$ να γίνεται

αρνητικός με θετική πιθανότητα.) Έτσι, για τις T_1, T_2 η κατανομή των συνιστωσών της μείξης στην (2) δίνεται από την ακόλουθη πρόταση.

Πρόταση 2. *Έστω* $U_1, \ldots, U_{n_1}, V_1, \ldots, V_{n_2} \overset{\text{iid}}{\sim} \mathcal{E}(1)$ *και* $s_i, t_j \geq 0, x > 0$, $i = 1, \ldots, n_1, j = 1, \ldots, n_2$. *Έστω* p *και* q *το πλήθος των διακεκριμένων* s_i *και διακεκριμένων* t_j, *αντίστοιχα, και έστω* (λ_a, γ_a), $a = 1, \ldots, p$, *και* (μ_b, δ_b), $b = 1, \ldots, q$, *οι αντίστροφοί τους με τις αντίστοιχες συχνότητές τους. Τότε,*

$$P\left\{ \frac{\sum_{i=1}^{n_1} s_i U_i}{\sum_{i=1}^{n_2} t_i V_i} \leqslant x \right\} = 1 - \sum_{a=1}^{p} \sum_{b=1}^{q} \sum_{j=1}^{\gamma_a} \sum_{k=1}^{\delta_b} \psi_{j,a}(\gamma, \lambda) \psi_{k,b}(\delta, x\mu) \Phi_{j,k}(\lambda_a, x\mu_b),$$

όπου $\gamma = (\gamma_1, \ldots, \gamma_p)$, $\lambda = (\lambda_1, \ldots, \lambda_p)$, $\delta = (\delta_1, \ldots, \delta_q)$, $\mu = (\mu_1, \ldots, \mu_q)$,

$$\psi_{j,a}(\gamma, \lambda) = (-\lambda_a)^{\gamma_a - j} \sum_{\substack{\sum_{i=1}^{p} \nu_i = \gamma_a - j \\ \nu_a = 0}} \prod_{\substack{i=1 \\ i \neq a}}^{p} \binom{\gamma_i + \nu_i - 1}{\nu_i} \frac{\lambda_i^{\gamma_i}}{(\lambda_i - \lambda_a)^{\gamma_i + \nu_i}},$$

$$\psi_{k,b}(\delta, \mu) = (-\mu_b)^{\delta_b - k} \sum_{\substack{\sum_{i=1}^{q} \nu_i = \delta_b - k \\ \nu_b = 0}} \prod_{\substack{i=1 \\ i \neq b}}^{q} \binom{\delta_i + \nu_i - 1}{\nu_i} \frac{\mu_i^{\delta_i}}{(\mu_i - \mu_b)^{\delta_i + \nu_i}}$$

και

$$\Phi_{j,k}(\lambda_a, \mu_b) = \sum_{i=0}^{j-1} \binom{i + k - 1}{i} \frac{\lambda_a^i \mu_b^k}{(\lambda_a + \mu_b)^{i+k}}.$$

Απόδειξη. Προφανώς, $\sum_{i=1}^{n_1} s_i U_i / \sum_{j=1}^{n_2} t_j V_j \overset{\text{d}}{=} \sum_{a=1}^{p} W_a / \sum_{b=1}^{q} Z_b$ όπου $W_1, \ldots, W_p$, $Z_1, \ldots, Z_q$ είναι ανεξάρτητες τυχαίες μεταβλητές Erlang, $W_a \sim \text{Erl}(\gamma_a, \lambda_a), Z_b \sim \text{Erl}(\delta_b, \mu_b)$. Από τους Amari and Misra (1997) γνωρίζουμε ότι

$$P\Big(\sum_{a=1}^{p} W_a \leqslant y \Big) = \sum_{a=1}^{p} \sum_{j=1}^{\gamma_a} \psi_{j,a}(\gamma, \lambda) G(y; j, \lambda_a) \tag{3}$$

όπου $G(y; j, \lambda_a) = 1 - \sum_{i=0}^{a-1} e^{-y\lambda_a} (y\lambda_a)^i / i!$, $x > 0$, είναι η συνάρτηση κατανομής της $\text{Erl}(\gamma_a, \lambda_a)$. Γράφοντας

$$P\left\{ \frac{\sum_{i=1}^{n_1} s_i U_i}{\sum_{i=1}^{n_2} t_i V_i} \leqslant x \right\} = P\Big\{ \sum_{a=1}^{p} W_a \leqslant x \sum_{b=1}^{q} Z_b \Big\},$$

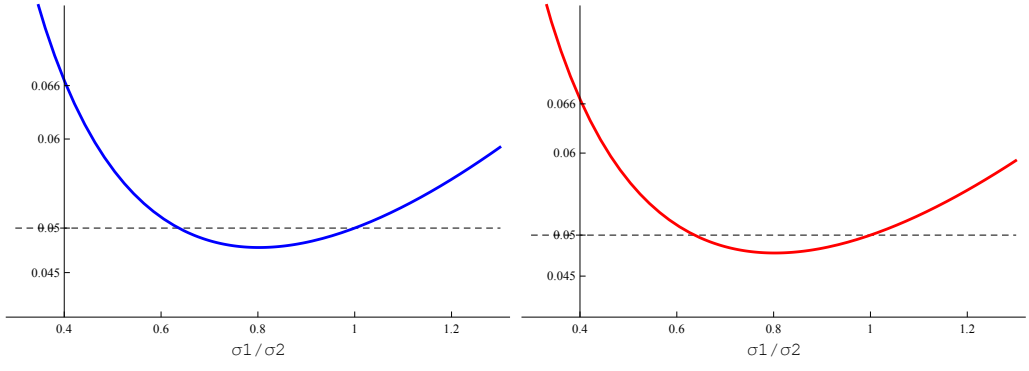
αντικαθιστώντας τις κατανομές των $\sum W_a$ και $\sum Z_b$ από την (3) και ολοκληρώνοντας παίρνουμε το αποτέλεσμα. □

Βασιζόμενοι στις ακριβείς εκφράσεις των συναρτήσεων κατανομής των T_1, T_2 μπορούμε να υπολογίσουμε αριθμητικά κατάλληλα ποσοστιαία σημεία για να κατασκευάσουμε τις περιοχές απόρριψης των ελέγχων καθώς και διαστήματα εμπιστοσύνης για τον λόγο σ_1/σ_2. Συμβολίζοντας με $T_{j,\alpha}$ το (άνω) α-ποσοστιαίο σημείο της T_j, $j = 1, 2$, ο απλούστερος έλεγχος μεγέθους α είναι εκείνος που απορρίπτει την υπόθεση $\sigma_1 = \sigma_2$ όταν $T_j < T_{j,1-\alpha/2}$ ή $T_j > T_{j,\alpha/2}$.

3. ΑΜΕΡΟΛΗΠΤΟΙ ΕΛΕΓΧΟΙ ΚΑΙ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Ως γνωστόν, σε κάθε $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για μία παράμετρο θ αντιστοιχεί ένας δίπλευρος έλεγχος μεγέθους α για αυτήν την παράμετρο και το αντίστροφο. Πιο συγκεκριμένα, το διάστημα εμπιστοσύνης περιέχει κάποια τιμή θ_0 αν και μόνον αν ο αντίστοιχος έλεγχος δεν απορρίπτει την υπόθεση $\theta = \theta_0$ (με εναλλακτική την $\theta \neq \theta_0$). Επίσης, ένας έλεγχος μεγέθους α για την παράμετρο θ είναι αμερόληπτος, δηλαδή η ισχύς του είναι τουλάχιστον α, αν και μόνον αν το αντίστοιχο διάστημα εμπιστοσύνης είναι αμερόληπτο, αν δηλαδή η πιθανότητα να συμπεριλάβει οποιαδήποτε τιμή θ πλην της πραγματικής είναι το πολύ $1-\alpha$.

Παρ΄ όλο που τα διαστήματα εμπιστοσύνης ίσων ουρών και οι αντίστοιχοι έλεγχοι (στους οποίους θα αναφερόμαστε στο εξής επίσης ως «ίσων ουρών») είναι βολικές διαδικασίες, όταν πρόκειται για παραμέτρους κλίμακος (όπως είναι εδώ το σ_1/σ_2 για την κατανομή των $T_1 = \hat{\sigma}_1/\hat{\sigma}_2$ και $T_2 = \tilde{\sigma}_1/\tilde{\sigma}_2$) έχουν το πρόβλημα της μεροληψίας. Στο Σχήμα 1 βλέπουμε τις καμπύλες ισχύος των ελέγχων μεγέθους $\alpha = 0.05$ ίσων ουρών που βασίζονται στις T_1 και T_2 όταν $n_1 = 2, n_2 = 3$: και στις δύο περιπτώσεις υπάρχουν τιμές σ_1/σ_2 για τις οποίες η ισχύς είναι μικρότερη του 5%.



Σχήμα 1: *Καμπύλη ισχύος των* $T_1,\ T_2$ *για* $\alpha = 0.05$.

Δεδομένου του ότι οι στατιστικές συναρτήσεις T_1 και T_2 και οι κατανομές τους έχουν την ίδια μορφή θα χρησιμοποιήσουμε στη συνέχεια και για τις δύο το σύμβολο T.

Λήμμα 1. *Ο έλεγχος που απορρίπτει την υπόθεση* $\sigma_1/\sigma_2 = 1$ *αν* $T < c_1$ *ή* $T > c_2$ *είναι αμερόληπτος μεγέθους* α *αν και μόνον αν*

$$F_T(c_2) - F_T(c_1) = 1 - \alpha \quad \textit{και} \quad c_1 f_T(c_1) = c_2 f_T(c_2), \tag{4}$$

όπου f_T *και* F_T *η πυκνότητα πιθανότητας και η συνάρτηση κατανομής της* T*, αντίστοιχα, όταν* $\sigma_1/\sigma_2 = 1$*.*

Η απόδειξη του λήμματος είναι απλή και βασίζεται στο γεγονός ότι η ισχύς του ελέγχου όταν $\sigma_1/\sigma_2 = \rho$ δίνεται από την διαφορά

$$F_T(\rho c_2) - F_T(\rho c_1) \tag{5}$$

και η συνθήκη για την ελαχιστοποίησή της στο $\rho = 1$ (έτσι ώστε η ισχύς να είναι μεγαλύτερη από α για $\rho \neq 1$) είναι η $c_2 f_T(c_2) - c_1 f_T(c_1) = 0$. Το λήμμα είναι γνωστό για την περίπτωση της διασποράς κανονικής κατανομής και της μέσης τιμής εκθετικής κατανομής (βλ. Lehmann, 1986).

Προκειμένου να καθορίσουμε τις κρίσιμες σταθερές c_1, c_2 για τον αμερόληπτο έλεγχο μεγέθους α θα πρέπει να λύσουμε το σύστημα στην (4). Εδώ έχουμε το εξής πρακτικό πρόβλημα. Δεδομένου του ότι η συνάρτηση κατανομής και η πυκνότητα πιθανότητας δίνονται από αθροίσματα $(n_1 + 1)(n_2 + 1)$ όρων με τον κάθε ένα από αυτούς να είναι ένα αρκετά πολύπλοκο άθροισμα, η επίλυση του συστήματος είναι πολύ χρονοβόρα. Γι΄ αυτό προτιμούμε να λύσουμε το σύστημα αριθμητικά αποφεύγοντας τον υπολογισμό των πυκνοτήτων πιθανότητας και σε αυτό μας βοηθάει η ακόλουθη παρατήρηση.

Γενικά, το διάστημα εμπιστοσύνης που αντιστοιχεί στον έλεγχο που απορρίπτει την $\sigma_1/\sigma_2 = 1$ αν $T < c_1$ ή $T > c_2$ είναι το $\left[T/c_2, T/c_1\right]$. Ο αναμενόμενος λόγος άκρων αυτού του διαστήματος είναι $(c_2/c_1)E_{\sigma_1=\sigma_2}(T)$ και η ελαχιστοποίησή του ισοδυναμεί προφανώς με ελαχιστοποιήση του λόγου c_2/c_1. Είναι εύκολο να διαπιστώσει κανείς ότι τα c_1, c_2 που αντιστοιχούν στο $100(1-\alpha)\%$ διάστημα εμπιστοσύνης ελαχίστου αναμενόμενου λόγου άκρων είναι ακριβώς η λύση του συστήματος στην (4). Επομένως, η επίλυση του συστήματος στην (4) ισοδυναμεί με την εύρεση των c_1, c_2 που ελαχιστοποιούν τον λόγο c_2/c_1 υπό τον περιορισμό $F_T(c_2) - F_T(c_1) = 1 - \alpha$. Επειδή δε

$$F_T(c_2) - F_T(c_1) = 1 - \alpha \Leftrightarrow c_2 = F_T^{-1}(1 - \alpha + F_T(c_1)),$$

το πρόβλημα ανάγεται στην εύρεση του c_1 που ελαχιστοποιεί την συνάρτηση

$$g(c_1) = F_T^{-1}(1 - \alpha + F_T(c_1))/c_1. \tag{6}$$

Για την ελαχιστοποίηση της g ως προς c_1 χρησιμοποιούμε την μέθοδο golden-section search (βλ. Cheney et al., 2008) την οποία περιγράφουμε ακολούθως.

Έστω g μία προς ελαχιστοποίηση συνάρτηση για την οποία γνωρίζουμε ότι το ελάχιστο αντιστοιχεί σε κάποιο σημείο του διαστήματος $[a, b]$ και ότι σε αυτό το διάστημα η g είναι πρώτα γνησίως φθίνουσα και μετά γνησίως αύξουσα. Θεωρούμε $x_1 < x_2 \in (a, b)$ και υπολογίζουμε τις τιμές $g(x_1), g(x_2)$.

- Αν $g(x_1) < g(x_2)$ τότε το ελάχιστο βρίσκεται στο διάστημα $[a, x_2]$ οπότε θέτουμε $b = x_2$, $x_2 = x_1$ και $x_1 = (\phi - 1)a + (2 - \phi)b$, όπου $\phi = (1 + \sqrt{5})/2 \approx 1.618$ είναι η σταθερά της χρυσής τομής, υπολογίζουμε τα νέο $g(x_1)$ και επαναλαμβάνουμε την σύγκριση.
- Αν $g(x_1) \geqslant g(x_2)$ τότε το ελάχιστο βρίσκεται στο διάστημα $[x_1, b]$ οπότε θέτουμε $a = x_1$, $x_1 = x_2$ και $x_2 = (2 - \phi)a + (\phi - 1)b$, υπολογίζουμε τα νέο $g(x_2)$ και επαναλαμβάνουμε την σύγκριση.

Η διαδικασία συνεχίζεται μέχρι το μήκος του διαστήματος $[a, b]$ να γίνει μικρότερο από κάποια προεπιλεγμένη σταθερά ϵ. Η επιλογή των κυρτών συνδυασμών $(\phi-1)a+(2-\phi)b$ και $(2-\phi)a+(\phi-1)b$ εξασφαλίζουν ότι έπειτα από k βήματα το διάστημα $[a, b]$ θα έχει μήκος ίσο με $(\phi - 1)^k$ φορές το μήκος του αρχικού διαστήματος.

Επιστρέφοντας στο πρόβλημα ελαχιστοποίησης της (6), λόγω της εξίσωσης $F_T(c_2) - F_T(c_1) = 1 - \alpha$ το ζητούμενο c_1 βρίσκεται υποχρεωτικά στο διάστημα $(0, F_T^{-1}(\alpha))$. Επομένως εφαρμόζοντας την παραπάνω μέθοδο μπορούμε να εντοπίσουμε το ελάχιστο χωρίς να χρειαστεί ο υπολογισμός της πυκνότητας πιθανότητας f_T στα διάφορα σημεία.

Η ισχύς των ελέγχων ίσων ουρών και των αμερόληπτων ελέγχων παρουσιάζεται στον Πίνακα 1 για επιλεγμένα μεγέθη δείγματος και τιμές του $\rho = \sigma_1/\sigma_2$. Από τα αποτελέσματα φαίνεται ότι οι έλεγχοι που βασίζονται στην στατιστική συνάρτηση $T_2 = \tilde{\sigma}_1/\tilde{\sigma}_2$ είναι ισχυρότεροι από εκείνους που βασίζονται στην $T_1 = \hat{\sigma}_1/\hat{\sigma}_2$ αλλά η διαφορά των ισχύων τους είναι ελάχιστη. Η μη αμεροληψία των ελέγχων ίσων ουρών φαίνεται μόνο στην περίπτωση των πολύ μικρών μεγεθών δείγματος ($n_1 = 2, n_2 = 3$) μια και όσο μεγαλώνουν τα μεγέθη δείγματος τόσο μειώνεται το διάστημα που περιέχει τα ρ για τα οποία η ισχύς είναι μικρότερη του α. Το διάστημα αυτό βρίσκεται είτε αριστερά είτε δεξιά της μονάδας, ανάλογα με το ποιο μέγεθος δείγματος είναι μεγαλύτερο.

4. ΑΡΙΘΜΗΤΙΚΟ ΠΑΡΑΔΕΙΓΜΑ

Σε αυτή την ενότητα παρουσιάζουμε τους ελέγχους και τα αμερόληπτα διαστήματα εμπιστοσύνης που περιγράψαμε στις προηγούμενες ενότητες μέσω ενός αριθμητικού παραδείγματος. Τα παρακάτω δεδομένα παριστάνουν αποδόσεις φοιτητών σε μία εργασία, σύμφωνα με την κλίμακα Wechsler Adult Intelligence. Οι φοιτητές χωρίστηκαν σε δύο ομάδες: «μη καπνιστές» εκείνοι που ισχυρίζονταν ότι δεν κάπνιζαν μαριχουάνα και «καπνιστές» όσοι ισχυρίζονταν ότι κάπνιζαν μαριχουάνα τακτικά. Τα δεδομένα είναι τα εξής:

Μη καπνιστές:	18,22,21,17,20,17,23,20,22,21
Καπνιστές:	16,20,14,21,20,18,13,15,17,21

Έστω X και Y τα τυχαία δείγματα μεγέθους $n_1 = n_2 = 10$ των μη καπνιστών και των καπνιστών αντίστοιχα. Υποθέτουμε ότι προέρχονται από τις κατανομές Laplace $\mathcal{L}(\mu_1, \sigma_1)$ και $\mathcal{L}(\mu_2, \sigma_2)$, αντίστοιχα.

Ο ΕΜΠ του σ_1 για τους μη καπνιστές είναι $\hat{\sigma}_1 = 1.7$ ενώ ο ΕΜΠ του σ_2 για τους καπνιστές είναι $\hat{\sigma}_2 = 2.5$. Συνεπώς, $T_1 = 1.7/2.5 = 0.68$. Οι ΒΓΑΕ είναι $\tilde{\sigma}_1 = 1.7754$, $\tilde{\sigma}_2 = 2.6320$, οπότε $T_2 = 1.7754/2.2037 \approx 0.675$. Χρησιμοποιώντας το `Mathematica v9.0` υπολογίσαμε τα ποσοστιαία σημεία $T_{1,0.975} = 0.394$, $T_{1,0.025} = 2.540$, $T_{2,0.975} = 0.394$, $T_{2,0.025} = 2.539$ (είναι διαφορετικά αλλά αυτό δεν φαίνεται στα πρώτα δεκαδικά ψηφία). Επειδή $T_{1,0.975} < T_1 < T_{1,0.025}$ και $T_{2,0.975} < T_2 < T_{2,0.025}$ συμπεραίνουμε ότι κανένας από τους δύο ελέγχους δεν απορρίπτει τη μηδενική υπόθεση ισότητας των παραμέτρων κλίμακας σε επίπεδο σημαντικότητας 5%.

Το 95% διάστημα εμπιστοσύνης ίσων ουρών για το λόγο σ_1/σ_2 που βασίζεται στην T_1 προκύπτει ως εξής: $[T_1/T_{1,0.025}, T_1/T_{1,0.975}] = [0.68/2.540, 0.68/0.394] = [0.268, 1.726]$. Ο λόγος των άκρων του παραπάνω διαστήματος είναι ίσος με $T_{1,0.025}/T_{1,0.975} = 2.540/0.394 = 6.447$. Το αντίστοιχο 95% αμερόληπτο διάστημα εμπιστοσύνης για το λόγο σ_1/σ_2 είναι: $[T_1/c_2, T_1/c_1] = [0.68/2.540, 0.68/0.394] = [0.268, 1.728]$. Κατά τον ίδιο τρόπο υπολογίζεται και το 95% διάστημα εμπιστοσύνης ίσων ουρών που βασίζεται στην T_2, καθώς και το αντίστοιχο αμερόληπτο διάστημα εμπιστοσύνης.

Ευχαριστίες: Αυτή η εργασία υποστηρίζεται εν μέρει από το Κέντρο Ερευνών του Πανεπιστημίου Πειραιώς.

ΑΝΑΦΟΡΕΣ

Amari, S.V., Misra, R.B. (1997). Closed-form expressions for distribution of sum of exponential random variables, *IEEE Transactions on Reliability*, **46**, 519-522.

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N. (2008). *A First Course in Order Statistics,* Classic Edition, SIAM, Philadelphia.

Cheney, W., Kincaid, D. (2008) *Numerical Mathematics and Computing, 6th Edition.* Brooks/Cole: Cengage Learning, Florence, Kentucky.

Govindarajulu, Z. (1966). Best linear unbiased estimates under symmetric censoring of the parameters of double exponential population, *Journal of the American Statistical Association*, **61**, 248-258.

Iliopoulos, G., Balakrishnan, N. (2009). Conditional independence of blocked ordered data, *Statistics and Probability Letters*,**79**, 1008-1015.

Iliopoulos, G., Balakrishnan, N. (2011). Exact likelihood inference for Laplace distribution based on type-II censored samples, *Journal of Statistical Planning and Inference*, **141**, 1224-1239.

Jasiulewicz, H., Kordecki, W. (2003). Convolutions of Erlang and of Pascal distributions with applications to reliability, *Demonstratio Mathematica*, **36**, 231-238.

Kotz, S., Kozubowski, T.J., Podgórski, K. (2001). *The Laplace Distribution and Generalizations,* Birkhäuser, Boston.

		$n_1 = 2$, $n_2 = 3$				$n_1 = 4$, $n_2 = 7$				$n_1 = 6$, $n_2 = 12$				$n_1 = 10$, $n_2 = 10$			
		EIO		AE		EIO		AE		EIO		AE		EIO		AE	
ρ	α	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
0.6	10%	.1030	.1030	.1198	.1198	.1620	.1624	.1793	.1797	.2202	.2207	.2394	.2398	.2876	.2879	.2876	.2879
	5%	.0512	.0512	.0607	.0607	.0891	.0893	.1009	.1011	.1307	.1309	.1447	.1450	.1860	.1862	.1860	.1862
	1%	.0102	.0102	.0123	.0123	.0206	.0206	.0244	.0244	.0353	.0354	.0406	.0407	.0614	.0615	.0614	.0615
0.7	10%	.0979	.0979	.1098	.1098	.1269	.1272	.1391	.1393	.1555	.1557	.1693	.1695	.1933	.1935	.1933	.1935
	5%	.0486	.0486	.0553	.0553	.0667	.0668	.0747	.0748	.0862	.0863	.0957	.0958	.1145	.1145	.1145	.1145
	1%	.0096	.0096	.0111	.0111	.0144	.0144	.0168	.0168	.0207	.0208	.0239	.0239	.0319	.0319	.0319	.0319
0.8	10%	.0964	.0964	.1039	.1039	.1078	.1079	.1155	.1156	.1188	.1189	.1275	.1276	.1368	.1368	.1368	.1369
	5%	.0478	.0478	.0521	.0521	.0547	.0548	.0597	.0598	.0621	.0621	.0679	.0679	.0746	.0747	.0746	.0747
	1%	.0095	.0095	.0105	.0105	.0112	.0112	.0127	.0127	.0135	.0135	.0153	.0153	.0178	.0178	.0178	.0178
0.9	10%	.0973	.0973	.1009	.1009	.0998	.0998	.1035	.1035	.1020	.1021	.1062	.1062	.1082	.1082	.1082	.1082
	5%	.0484	.0484	.0505	.0505	.0498	.0498	.0522	.0522	.0513	.0513	.0540	.0540	.0554	.0554	.0554	.0554
	1%	.0096	.0096	.0101	.0101	.0099	.0099	.0106	.0106	.0103	.0103	.0112	.0112	.0117	.0117	.0117	.0117
1.1	10%	.0964	.1041	.1007	.1007	.1063	.1063	.1029	.1029	.1090	.1091	.1052	.1052	.1067	.1067	.1067	.1067
	5%	.0524	.0524	.0504	.0504	.0541	.0541	.0519	.0519	.0560	.0560	.0534	.0534	.0544	.0544	.0544	.0544
	1%	.0106	.0106	.0101	.0101	.0112	.0112	.0105	.0105	.0119	.0119	.0110	.0110	.0113	.0114	.0113	.0114
1.2	10%	.1093	.1093	.1027	.1027	.1173	.1173	.1107	.1108	.1266	.1266	.1192	.1192	.1246	.1246	.1246	.1246
	5%	.0555	.0555	.0515	.0515	.0614	.0614	.0569	.0570	.0679	.0680	.0628	.0628	.0663	.0664	.0663	.0664
	1%	.0113	.0113	.0103	.0103	.0135	.0135	.0120	.0120	.0158	.0158	.0141	.0141	.0151	.0151	.0151	.0151
2	10%	.1704	.1704	.1431	.1431	.2814	.2820	.2577	.2584	.3909	.3915	.3672	.3679	.4289	.4293	.4289	.4293
	5%	.0938	.0938	.0756	.0756	.1834	.1838	.1640	.1644	.2794	.2799	.2582	.2587	.3051	.3055	.3051	.3055
	1%	.0212	.0212	.0162	.0162	.0618	.0619	.0524	.0525	.1167	.1170	.1040	.1042	.1222	.1224	.1222	.1224
5	10%	.4071	.4071	.3487	.3487	.7577	.7593	.7354	.7372	.9122	.9129	.9030	.9037	.9602	.9604	.9602	.9604
	5%	.2788	.2788	.2265	.2266	.6601	.6618	.6329	.6346	.8619	.8627	.8487	.8496	.9243	.9247	.9243	.9247
	1%	.0910	.0910	.0668	.0668	.4290	.4304	.3974	.3987	.7083	.7095	.6867	.6880	.7842	.7849	.7842	.7849
10	10%	.6233	.6233	.5671	.5671	.9418	.9426	.9343	.9352	.9925	.9926	.9914	.9916	.9991	.9991	.9991	.9991
	5%	.4976	.4976	.4340	.4340	.9061	.9072	.8947	.8959	.9861	.9863	.9842	.9844	.9977	.9977	.9977	.9977
	1%	.2305	.2305	.1809	.1809	.7836	.7854	.7613	.7631	.9577	.9581	.9525	.9530	.9868	.9869	.9868	.9869

Πίνακας 1: *Ισχύς ελέγχων ίσων ουρών και αμερόληπτων ελέγχων για διάφορα μεγέθη δείγματος.*

Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 30ου Πανελληνίου Συνεδρίου Στατιστικής (2017), σελ.80-95

ΕΛΕΓΧΟΣ ΗΜΕΡΗΣΙΑΣ ΧΡΗΣΗΣ ΤΟΥ ΥΠΟΛΟΓΙΣΤΗ ΔΙΑ ΜΕΣΩ ΔΥΝΑΜΙΚΗΣ ΤΗΣ ΠΛΗΚΤΡΟΛΟΓΗΣΗΣ

Τσιμπερίδης Ιωάννης, Καράκος Αλέξανδρος
Δημοκρίτειο Πανεπιστήμιο Θράκης
itsimper@ee.duth.gr, karakos@ee.duth.gr

ΠΕΡΙΛΗΨΗ

Το Διαδίκτυο αποτελεί ένα εργαλείο, με το οποίο σήμερα συνδέονται δισεκατομμύρια άνθρωποι με σκοπό την επικοινωνία, τη διασκέδαση, την εργασία και τη μόρφωση. Παράλληλα όμως με τα οφέλη που αποκομίζονται, ελλοχεύουν και κίνδυνοι που μπορεί να βλάψουν τους χρήστες οικονομικά, ηθικά και κοινωνικά. Πολλοί από αυτούς τους κινδύνους προέρχονται από κακόβουλους χρήστες, των οποίων ένα από τα σημαντικότερα όπλα είναι η ανωνυμία που μπορούν να διατηρήσουν στο χώρο του Διαδικτύου. Στην παρούσα εργασία προτείνεται μία μέθοδος, η οποία αξιοποιεί τα δεδομένα που παράγονται από τον τρόπο που πληκτρολογούν οι χρήστες. Σκοπός είναι η εύρεση κάποιων χαρακτηριστικών τους, έτσι ώστε να αρθεί η πλήρης ανωνυμία και να είναι εφικτή η προειδοποίηση των ανυποψίαστων χρηστών για την πιθανότητα ο συνομιλητής τους να έχει παραποιήσει τα χαρακτηριστικά του. Το προτεινόμενο σύστημα υλοποιήθηκε με τεχνητά νευρωνικά δίκτυα και το ποσοστό ορθής πρόβλεψης ανήλθε στο 75%, υπερβαίνοντας κατά πολύ αυτό της τυχαίας πρόβλεψης.

Λέξεις Κλειδιά: Δυναμική της Πληκτρολόγησης, Κατηγοριοποίηση Χρηστών, Μάθηση Μηχανής, Εγκληματολογικά Πειστήρια.

1. ΕΙΣΑΓΩΓΗ

Στο Διαδίκτυο σήμερα εκτυλίσσεται ένα μεγάλο μέρος των καθημερινών ανθρώπινων δραστηριοτήτων, όπως της εργασίας, της διασκέδασης, της εκπαίδευσης και της κοινωνικής δικτύωσης. Εκτός από την αύξηση στο πλήθος των υπηρεσιών που προσφέρονται, αλλά και τη διεύρυνση της ποικιλίας τους, το πλήθος των χρηστών που συνδέονται αυξάνεται με ρυθμούς πολύ μεγαλύτερους από αυτούς του παγκόσμιου πληθυσμού, με τους αριθμούς να μιλάνε για 3,5 δισεκατομμύρια χρήστες του Διαδικτύου.

Παράλληλα, αύξηση της ποικιλομορφίας παρουσιάζεται και στους τρόπους με τους οποίους οι χρήστες μπορούν να επικοινωνήσουν μεταξύ τους. Έτσι, για παράδειγμα, υπάρχει η δυνατότητα επικοινωνίας με άμεσα ή έμμεσα γραπτά μηνύματα, με μετάδοση φωνής, με βίντεο, με διαμοιραζόμενα μέσα, ή και με άλλους

τρόπους. Παρότι όμως τα μέσα επικοινωνίας γίνονται περισσότερα, αλλά και πιο εξελιγμένα, και παρότι οι ταχύτητες σύνδεσης γίνονται υψηλότερες, με τεχνολογίες που προσφέρουν διάφορους τρόπους πρόσβασης, ενσύρματης ή ασύρματης, μελέτες και στατιστικές (Internet live stats, n.d.) δείχνουν ότι το γραπτό κείμενο παραμένει το κύριο μέσο επικοινωνίας μεταξύ των χρηστών του Διαδικτύου. Η κύρια συσκευή εισόδου, με την οποία παράγεται κείμενο, είναι το πληκτρολόγιο QWERTY. Αν και προτάθηκαν διάφορες διατάξεις πληκτρολογίου, μερικές από τις οποίες συναντώνται και σήμερα σε διάφορες ηλεκτρονικές συσκευές, το πληκτρολόγιο QWERTY κατέχει τη μερίδα του λέοντος στις συσκευές σύνταξης κειμένου. Μάλιστα, εκτός από την παραδοσιακή μορφή του, ως αναπόσπαστο κομμάτι των desktops και των laptops, σήμερα έχει και εικονική μορφή στις πιο «σύγχρονες» και περισσότερο φορητές συσκευές, τα tablets και τα smartphones.

Εκτός όμως από την πληθώρα των υπηρεσιών του Διαδικτύου, παρουσιάζεται και μία ιδιομορφία στον τρόπο επικοινωνίας μεταξύ των χρηστών. Αυτή είναι η δυνατότητα κάποιου να παραμείνει ανώνυμος, είτε αποκρύπτοντας τα δηλωτικά χαρακτηριστικά του, είτε υιοθετώντας κάποια αναληθή. Η ανωνυμία αυτή είναι ως ένα σημείο επιθυμητή αφού παρέχει κάποια αίσθηση ελευθερίας στους χρήστες και κάποια διασφάλιση ότι δεν θα διαρρεύσουν προσωπικά ή ευαίσθητα δεδομένα τους. Πολλές φορές όμως αποτελεί πρόβλημα ή πηγή προβλημάτων.

Για παράδειγμα, όταν ο χρήστης αποκρύπτει την ταυτότητά του, δεν επωφελείται από υπηρεσίες του Διαδικτύου που εστιάζουν στα χαρακτηριστικά του και που μπορεί να του πρότειναν να επισκεφτεί συγκεκριμένους Δικτυακούς Τόπους του ενδιαφέροντός του, ή να συμμετέχει σε συζητήσεις των προτιμήσεών του, ή να ενταχθεί σε ομάδες με μέλη που έχουν κοινά χαρακτηριστικά με αυτόν.

Σημαντικότερο όμως είναι ότι η ανωνυμία τροποποιεί τη συμπεριφορά ενός χρήστη (Suler, 2004), αφού αίρονται κάποιες αναστολές του και πιθανόν, πέρα από την επιθυμητή απελευθέρωση της ευπρεπούς έκφρασης, να οδηγείται και σε παράνομες ενέργειες, τις οποίες δεν θα διέπραττε εάν η ταυτότητά του ήταν γνωστή. Επιπλέον, η ανωνυμία αποτελεί το μεγαλύτερο πλεονέκτημα των παράνομων χρηστών, οι οποίοι δίνοντας ψεύτικες πληροφορίες για την ταυτότητά τους, προσπαθούν να αποκτήσουν την εμπιστοσύνη ανυποψίαστων χρηστών με σκοπό την εκμετάλλευση ή βλάβη τους. Οικονομικές απάτες, αποπλάνηση ανηλίκων, σχολικός εκφοβισμός, συκοφαντία, διανομή παιδικής πορνογραφίας, απειλές, διανομή ιών υπολογιστών, κ.α., είναι ορισμένες από τις κακόβουλες ενέργειες που βασίζονται στην ανωνυμία.

Άρα, η αποκάλυψη κάποιων εγγενών ή επίκτητων χαρακτηριστικών ενός χρήστη, όπως το φύλο, η ηλικία, το επικρατές χέρι, το μορφωτικό επίπεδο, η μητρική γλώσσα, κ.α., τα οποία ηθελημένα ή από αμέλεια δεν δήλωσε κατά τη συνομιλία του με άλλον χρήστη ή κατά τη χρήση κάποιων Διαδικτυακών υπηρεσιών, θα ήταν ωφέλιμη για την καλύτερη εκμετάλλευση των δυνατοτήτων του Διαδικτύου, για την ενημέρωση ανυποψίαστων χρηστών για ενδεχόμενους κινδύνους, αλλά και για την παροχή χρήσιμων πληροφοριών, εγκληματολογικού ενδιαφέροντος, στις περιπτώσεις όπου κάποιο ηλεκτρονικό αδίκημα έχει διαπραχθεί.

Στην παρούσα εργασία επιχειρείται η ταξινόμηση χρηστών με σκοπό τον εντοπισμό κάποιων χαρακτηριστικών τους. Αυτό επιτυγχάνεται με τη βοήθεια της δυναμικής της πληκτρολόγησης (keystroke dynamics), δηλαδή, του τρόπου με τον οποίο ένας χρήστης πληκτρολογεί. Ως παράμετροι χρησιμοποιούνται διάρκειες πατήματος πλήκτρου (keystroke durations) και λανθάνοντες χρόνοι διγράμματος DDDL (down-down digram latency), ενώ το χαρακτηριστικό που αναζητείται είναι η ημερήσια χρήση της ηλεκτρονικής τους συσκευής. Η γνώση του πλήθους των ωρών ανά ημέρα που ξοδεύει ένας χρήστης σε υπολογιστή δεν μπορεί να αποτελέσει ασφαλές συμπέρασμα για τα χαρακτηριστικά του, αλλά μπορεί να οδηγήσει σε κάποιες υποθέσεις. Για παράδειγμα, να καταδείξει την ηλικία του, αφού σύμφωνα με στατιστικές μελέτες (Daily computer usage in Great Britain by age 2006-2015, n.d.) αυτά τα δύο μεγέθη συνδέονται αντιστρόφως ανάλογα, το φύλο του, καθώς οι άντρες καταναλώνουν περισσότερο χρόνο στον υπολογιστή από ότι οι γυναίκες (Winn & Heeter, 2009), και το ετήσιο εισόδημά του, εφόσον φαίνεται ότι τα υψηλότερα εισοδήματα αντιστοιχίζονται σε λιγότερες ώρες μπροστά σε υπολογιστή (Hofferth et al., 2013). Ακόμα, η γνώση αυτή μπορεί να βοηθήσει στον αποκλεισμό υπόπτων στην περίπτωση εγκληματολογικής έρευνας, αφού μάλλον θα ήταν απίθανο ένα άτομο με ελάχιστες ώρες ενασχόλησης να διαπράξει μία κυβερνοεπίθεση, ή κάτι παραπλήσιο.

2. ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Για την επίλυση του προβλήματος της άρσης της πλήρους ανωνυμίας στο Διαδίκτυο έχουν προταθεί διάφορες μέθοδοι. Μία από αυτές είναι η εκμετάλλευση πληροφοριών από φωτογραφίες προσώπου που μπορεί ένας χρήστης να έχει στο προφίλ του. Έτσι οι Alrashed και Berbar (2013) σχεδίασαν σύστημα βασιζόμενο σε ταξινομητή μηχανής υποστήριξης διανυσμάτων (support vector machine, SVM) και εξετάζοντας 1982 φωτογραφίες πέτυχαν ένα ποσοστό ορθής πρόβλεψης φύλου κοντά στο 99,5%. Για την ακρίβεια των αποτελεσμάτων τους χρησιμοποίησαν τα στατιστικά μεγέθη ευαισθησία (sensitivity), το οποίο υπολογίζει το ποσοστό των αληθώς θετικών (true positive rate) αποτελεσμάτων, και ειδικότητα (specificity), το οποίο υπολογίζει το ποσοστό των αληθώς αρνητικών (true negative rate) αποτελεσμάτων. Επίσης, οι Damayanti και Rachmad (2016) σχεδίασαν παρόμοιο σύστημα, με την ακρίβειά του να κυμαίνεται από 74% έως 92%. Στο πεδίο της αναζήτησης της ηλικίας ενός χρήστη από φωτογραφίες προσώπου, οι Hewahi κ.ά. (2010) όρισαν 4 ηλικιακές ομάδες, επέλεξαν 68 σημεία πάνω σε μία εικόνα προσώπου και πέτυχαν ποσοστό ορθής πρόβλεψης 78,4%. Ο Choobeh (2012) με τον ίδιο αριθμό σημείων, επέλεξε 130 παραμέτρους, τις οποίες εφάρμοσε σε καθένα από τα νευρωνικά δίκτυα που ενεπλάκησαν στη διαδικασία πετυχαίνοντας ένα μέσο απόλυτο σφάλμα (mean absolute error) της ακριβούς ηλικίας του χρήστη μεταξύ 4,85 και 5,85 έτη. Ενώ, οι Yi κ.ά. (2014), χρησιμοποιώντας ένα συνελικτικό νευρωνικό δίκτυο (convolutional neural network, CNN), μείωσαν περαιτέρω το μέσο απόλυτο σφάλμα στα 3,63 έτη, βελτιώνοντας ταυτόχρονα και την ταχύτητα του συστήματος.

Μία άλλη μέθοδος που αφορά την εξεύρεση κάποιων χαρακτηριστικών των χρηστών, είναι εξέταση των κειμένων που παράγουν. Όπως η προσπάθεια των

Mukherjee και Liu (2010) για να κατατάξουν τους συγγραφείς των blogs σύμφωνα με το φύλο τους. Το σύστημά τους υλοποιήθηκε και με ταξινομητή Naïve Bayes, αλλά και με SVM, επιτυγχάνοντας ποσοστό ορθής πρόβλεψης της τάξης του 88%. Για την επιλογή των κατάλληλων παραμέτρων (features selection) που θα τους οδηγούσαν στα επιθυμητά αποτελέσματα, οι ερευνητές χρησιμοποίησαν το στατιστικό χ^2 (chi square statistic) που μετρά την έλλειψη ανεξαρτησίας μεταξύ μιας παραμέτρου και μιας κλάσης. Επίσης, οι Cheng κ.ά. (2011) προσπάθησαν να λύσουν το ίδιο πρόβλημα με δεδομένα προερχόμενα από μία συλλογή κειμένων από ομάδες συζητήσεων του Reuters και μια συλλογή από e-mails των εργαζομένων της εταιρίας Enron. Μελέτησαν 545 παραμέτρους, εκπαίδευσαν τρεις ταξινομητές και τα αποτελέσματα έδειξαν τον SVM ως πιο αποδοτικό, το σύστημα να βελτιώνεται όσο αύξανε το σύνολο εκπαίδευσης (training set) και όσο ελέγχονταν μεγαλύτερα κείμενα, με ένα μέγιστο ποσοστό επιτυχίας της τάξης του 85%. Μάλιστα, σε μια διαδικασία μείωσης του πλήθους των παραμέτρων, έτσι ώστε να μπορέσουν να κάνουν το σύστημά τους ταχύτερο, οι ερευνητές αξιολόγησαν τη σημαντικότητα των παραμέτρων χρησιμοποιώντας τη στατιστική δοκιμή t-test. Θέτοντας το επίπεδο σημαντικότητας στο 5%, κατέληξαν σε 157 παραμέτρους, με τη χρήση των οποίων το σύστημα γινόταν 3 φορές ταχύτερο, με απώλεια 3% στην ακρίβεια.

Η αναζήτηση του φύλου και της ηλικίας του δημιουργού ενός blog ήταν το αντικείμενο μελέτης της εργασίας των Schler κ.ά. (2006). Οι ερευνητές όρισαν τρεις κλάσεις, την 10's που αντιστοιχούσε στις ηλικίες 13-17, την 20's που αντιστοιχούσε στις 23-27, και στην 30's που αντιστοιχούσε στις 33-42. Για την ταξινόμηση επιστρατεύτηκε ο αλγόριθμος Multi-Class Real Winnow και τα τελικά αποτελέσματα έδειξαν ότι η ηλικιακή ομάδα μπορούσε να προβλεφθεί ορθώς με ένα ποσοστό της τάξης του 73%, ενώ το φύλο κοντά στο 80%.

Τέλος, τα χαρακτηριστικά των χρηστών αναζητήθηκαν με παραμέτρους που προέκυψαν από τη συμπεριφορά τους σε κοινωνικά δίκτυα. Για παράδειγμα, στην εργασία των Rao κ.ά. (2010) επιχειρείτε η αναγνώριση χαρακτηριστικών των χρηστών του Twitter. Μεταξύ των πεδίων στα οποία αναζητήθηκαν παράμετροι ήταν η δομή του κοινωνικού τους δικτύου και η κοινωνική συμπεριφορά τους. Οι χρήστες χωρίστηκαν σε δύο ηλικιακές ομάδες και οι ερευνητές δοκίμασαν το σύστημά τους για τις παραμέτρους καθενός από τα πεδία που εντόπισαν, πετυχαίνοντας ποσοστό ορθής πρόβλεψης κοντά στο 74%.

Ωστόσο, όλες οι παραπάνω μέθοδοι αντιμετωπίζουν κάποιους περιορισμούς στη γενίκευσή τους. Ο λόγος είναι ότι απαιτούνται συγκεκριμένα δεδομένα (π.χ. φωτογραφίες προσώπου) από πλευράς χρήστη, ή υπάρχει εξάρτηση από συγκεκριμένη ομιλούμενη γλώσσα, αφού οι παράμετροι προέρχονται από συγκεκριμένες φράσεις, λέξεις και N-γράμματα, ή θεωρείται δεδομένη μια φυσιολογική συμπεριφορά του χρήστη, ο οποίος πρέπει να διατηρεί λογαριασμό σε κάποιο κοινωνικό δίκτυο. Στην παρούσα εργασία προτείνεται ο εντοπισμός κάποιων εγγενών ή επίκτητων χαρακτηριστικών χρήστη δια μέσω της δυναμικής της πληκτρολόγησης, που ορίζεται ως η λεπτομερής χρονική καταγραφή των ενεργειών του στο πληκτρολόγιο, δηλαδή το πότε πίεσε και το πότε απελευθέρωσε κάθε πλήκτρο. Η μέθοδος αυτή, εκτός του ότι δεν απαιτεί παρά τα πιο απλά δεδομένα που

παράγει ο χρήστης και εκτός του ότι είναι ανεξάρτητη από την ομιλούμενη γλώσσα που χρησιμοποιεί, δεν είναι παρεμβατική, με την έννοια ότι δεν είναι αναγκαία η ανάγνωση των κειμένων του, η παρακολούθηση της συμπεριφοράς του στο Διαδίκτυο και η αναζήτηση των συνδέσεων του κοινωνικού του δικτύου.

Η δυναμική της πληκτρολόγησης χρησιμοποιήθηκε ως επί το πλείστον στην αυθεντικοποίηση χρηστών, με σκοπό την αντικατάσταση του παραδοσιακού σχήματος με τη χρήση του κωδικού πρόσβασης. Όπως για παράδειγμα στην εργασία των Patil και Renke (2016), οι οποίοι κατέγραψαν χρήστες μιας εφαρμογής, αποθηκεύοντας τα δεδομένα σε βάση δεδομένων, και τα οποία στη συνέχεια τα χρησιμοποιούσαν για να επαληθεύσουν εάν ο χρήστης που πληκτρολογεί είναι αυτός που ισχυρίζεται ότι είναι. Επίσης, οι Wankhede και Verma (2014) πέτυχαν ποσοστό εσφαλμένης απόρριψης (false rejection rate, FRR) 4,8 % και ποσοστό εσφαλμένης αποδοχής (false acceptance rate, FAR) 3,1% στο σύστημα αυθεντικοποίησής τους, που υλοποιήθηκε με πολυστρωματικό perceptron (multi-layer perceptron, MLP). Μάλιστα, με σκοπό την απομάκρυνση δεδομένων που αύξαναν το βαθμό ασυνέπειας των χρηστών στην πληκτρολόγηση, χρησιμοποίησαν το στατιστικό μέγεθος Z-score, το οποίο υπολογίζεται με τη βοήθεια της μέσης τιμής και της τυπικής απόκλισης των τιμών κάθε παραμέτρου.

Από τη δυναμική της πληκτρολόγησης μπορούν να εξαχθούν πολλές παράμετροι, κάθε μία από τις οποίες περικλείει μικρή ποσότητα πληροφορίας, ο συνδυασμός τους όμως είναι ικανός να δώσει ικανοποιητικά αποτελέσματα, όπως τουλάχιστον φάνηκε από σχετικές μελέτες,. Έτσι, χρησιμοποιείται η διάρκεια πατήματος πλήκτρου (keystroke duration) που ορίζεται ως ο χρόνος που πέρασε από τη στιγμή που ένα πλήκτρο πατήθηκε, μέχρι τη στιγμή που ελευθερώθηκε. Μια άλλη παράμετρος είναι ο λανθάνων χρόνος διγράμματος (digram latency) που ορίζεται ως ο χρόνος που χρειάστηκε ένας χρήστης για να χρησιμοποιήσει δύο συνεχόμενα πλήκτρα. Το μέγεθος αυτό μπορεί να εκφραστεί με τέσσερις διαφορετικούς τρόπους (Hosseinzadeh & Krishnan, 2008), που προκύπτουν από τους συνδυασμούς πίεσης και απελευθέρωσης των δύο πλήκτρων, και συγκεκριμένα είναι τα down-down digram latency (DDDL), up-up digram latency (UUDL), down-up digram latency (DUDL) και up-down digram latency (UDDL). Με παρόμοιους τρόπους ορίζονται και ο λανθάνων χρόνος τριγράμματος (trigram latency), ο λανθάνων χρόνος τετραγράμματος (tetragram latency) και γενικά ο λανθάνων χρόνος N-γράμματος (N-gram latency) (Zhao, 2006).

Εκτός όμως από τις παραμέτρους που σχετίζονται με το χρόνο, στις μελέτες που διεξήχθησαν και αφορούσαν τη δυναμική της πληκτρολόγησης, εντάσσονται και άλλες παράμετροι που δεν σχετίζονται με τη λεπτομερή καταγραφή των χρόνων των συμβάντων. Ως τέτοιες λογίζονται η ταχύτητα πληκτρολόγησης (λέξεις ανά λεπτό), η συχνότητα λαθών κατά την πληκτρολόγηση, ο τρόπος διόρθωσης λαθών, το ποσοστό χρήσης πλήκτρων που συναντώνται παραπάνω από μία φορές στο πληκτρολόγιο (όπως το “Shift” (Bartlow & Cukic, 2006), το “Ctrl”, το “Alt”, το “Enter”, κτλ) (Kumar et al, 2014), η ώρα της ημέρας που κάποιος χρήστης επιλέγει να πληκτρολογήσει, οι εφαρμογές στις οποίες πληκτρολογεί και γενικώς η συχνότητα χρήσης του πληκτρολογίου.

3. ΜΕΘΟΔΟΛΟΓΙΑ

Η μέθοδος που ακολουθήθηκε αποτελείται από τρία διακριτά στάδια. Στο πρώτο στάδιο έγινε η λήψη των απαραίτητων δεδομένων δυναμικής της πληκτρολόγησης από εθελοντές χρήστες. Στο δεύτερο στάδιο έγινε η επιλογή των παραμέτρων που χρησιμοποιήθηκαν για την κατηγοριοποίηση των χρηστών. Τέλος, στο τρίτο στάδιο έγινε η επιλογή που κατάλληλου ταξινομητή, έτσι ώστε να προκύψει το σύστημα με τη βέλτιστη απόδοση, σε ότι αφορά το ποσοστό ορθής πρόβλεψης, την ταχύτητα λειτουργίας και την σταθερότητα στην εξαγωγή αποτελεσμάτων.

3.1 Λήψη Δεδομένων

Τα δεδομένα που λαμβάνονται από τη δυναμική της πληκτρολόγησης μπορεί να προέρχονται από καθορισμένο κείμενο (fixed) ή από ελεύθερο κείμενο (free text). Ως ελεύθερο κείμενο νοείται ένα συγκεκριμένο κείμενο που έχει δοθεί σε έναν εθελοντή χρήστη να πληκτρολογήσει ενώ βρίσκεται σε κατάσταση καταγραφής της πληκτρολόγησής του. Συνήθως το καθορισμένο κείμενο πληκτρολογείται σε κάποιο κλειστό περιβάλλον. Ως ελεύθερο κείμενο νοείται το κείμενο που πληκτρολογεί ο εθελοντής χρήστης κατά βούληση ενώ βρίσκεται σε κατάσταση καταγραφής της πληκτρολόγησής του. Καταγραφή πληκτρολόγησης ελεύθερου κειμένου μπορεί να διεξάγεται σε κλειστό περιβάλλον ή κατά τη διάρκεια καθημερινής χρήσης του υπολογιστή, όπου ο χρήστης χρησιμοποιεί τις εφαρμογές που επιθυμεί και πληκτρολογεί τις χρονικές στιγμές που επιθυμεί.

Στη συγκεκριμένη έρευνα, για λόγους που έχουν να κάνουν με την όσο το δυνατόν πλησιέστερη αναπαράσταση των πραγματικών συνθηκών, αλλά και για την καταγραφή παραμέτρων που δεν βρίσκονταν στον αρχικό σχεδιασμό, επιλέχθηκε η λήψη δεδομένων από ελεύθερο κείμενο. Όμως, τα διαθέσιμα σύνολα δεδομένων (datasets) της δυναμικής της πληκτρολόγησης στο Διαδίκτυο είναι ελάχιστα, με αυτά που προέρχονται από ελεύθερο κείμενο να είναι η μειονότητα. Μάλιστα, δεν φαίνεται να υπάρχει ούτε ένα σύνολο δεδομένων ελεύθερου κειμένου που να προήλθε από την καταγραφή μεγάλου κειμένου, τουλάχιστον σύμφωνα με όσα είναι γνωστά,. Τα όσα είναι διαθέσιμα περιορίζονται στην αποτύπωση ορισμένων λέξεων ή φράσεων, με τον λόγο να είναι προφανής και να μην είναι άλλος από το ότι αυτά τα δεδομένα μπορούν να αποκαλύψουν κωδικούς πρόσβασης, αριθμούς πιστωτικών καρτών, προσωπικά μηνύματα, και άλλες ευαίσθητες πληροφορίες του καταγεγραμμένου εθελοντή.

Για το λόγο αυτό αποφασίστηκε να δημιουργηθεί ένα νέο σύνολο δεδομένων για τις ανάγκες αυτής της έρευνας. Για την εκπλήρωση αυτού του στόχου σχεδιάστηκε λογισμικό καταγραφής πληκτρολόγησης (keylogger), με το όνομα «IRecU», το οποίο εγκαταστάθηκε στις συσκευές των εθελοντών. Κατά την πρώτη είσοδο του χρήστη στο «IRecU» του ζητούταν να συμπληρώσει μία φόρμα με ορισμένα χαρακτηριστικά του, ανάμεσα στα οποία ήταν η κατά μέσο όρο ημερήσια χρήση της ηλεκτρονικής συσκευής του. Οι επιλογές που δόθηκαν στους χρήστες ήταν 5, οι «0-1 ώρες», «1-2 ώρες», «2-4 ώρες», «4-6 ώρες» και «6+ ώρες», με συνέπεια να δημιουργηθούν 5

κλάσεις. Μετά το πέρας της διαδικασίας καταγραφής, η οποία διήρκεσε από 20/02/2014 έως και 27/12/2014, συλλέχθηκαν 248 αρχεία από 75 εθελοντές χρήστες. Κάθε αρχείο περιέχει δεδομένα από 2.800 έως 4.500 πατήματα πλήκτρων, καταγεγραμμένα σε εγγραφές της μορφής:

78,#2014-06-20#,34680537,"dn"
78,#2014-06-20#,34680657,"up"
65,#2014-06-20#,34680687,"dn"
73,#2014-06-20#,34680787,"dn"
65,#2014-06-20#,34680797,"up"
73,#2014-06-20#,34680887,"up"

Τα πεδία σε κάθε εγγραφή χωρίζονται με κόμμα (,). Στο πρώτο πεδίο δηλώνεται το virtual key code του πλήκτρου στο οποίο έγινε η ενέργεια, σε δεκαδική μορφή. Στο δεύτερο πεδίο, ανάμεσα στα σύμβολα της δίεσης (#), δηλώνεται η ημερομηνία που έγινε η ενέργεια πληκτρολόγησης. Στο τρίτο πεδίο δηλώνονται τα ms που πέρασαν από την αρχή της συγκεκριμένης ημέρας, τη στιγμή που έγινε η ενέργεια. Τέλος, στο τέταρτο πεδίο δηλώνεται το είδος της ενέργειας, με "dn" να αντιστοιχεί στην πίεση πλήκτρου και με "up" στην απελευθέρωση πλήκτρου.

Το διαθέσιμο πλήθος αρχείων ανά κλάση και το ποσοστό τους επί του συνόλου, παρουσιάζεται στον Πίνακα 1.

Πίνακας 1. *Πλήθος και ποσοστό αρχείων ανά κλάση*

	Πλήθος	**Ποσοστό**
Αρχεία 0-1 ωρών	23	9,3%
Αρχεία 1-2 ωρών	46	18,5%
Αρχεία 2-4 ωρών	54	21,8%
Αρχεία 4-6 ωρών	38	15,3%
Αρχεία 6 και άνω ωρών	87	35,1%
Συνολικά αρχεία	248	100,0%

Αν και το σύνολο δεδομένων δεν είναι ισορροπημένο, η αντιπροσώπευση κάθε κλάσης είναι επαρκής, με συνέπεια να θεωρούνται αξιόπιστα τα αποτελέσματα.

3.2 Εξαγωγή Παραμέτρων

Όπως προαναφέρθηκε, η δυναμική της πληκτρολόγησης συνοδεύεται από μεγάλο πλήθος παραμέτρων, με τις περισσότερες έρευνες να χρησιμοποιούν τις διάρκειες πατήματος πλήκτρου και τους λανθάνοντες χρόνους διγράμματος. Κάποιοι (Douhou & Magnus, 2009) ισχυρίζονται ότι η χρήση των keystroke durations φέρνει καλύτερα αποτελέσματα, ενώ κάποιοι άλλοι (Hassan et al., 2013) ότι οι λανθάνοντες χρόνοι διγράμματος είναι προτιμότεροι. Για το λόγο αυτό, όπως ήδη ειπώθηκε, σε αυτή την

εργασία χρησιμοποιούνται και τα δύο είδη παραμέτρων. Μάλιστα, για την αποφυγή αρνητικών τιμών, χρησιμοποιούνται οι λανθάνοντες χρόνοι διγράμματος DDDL.

Όμως, θεωρώντας ένα πληκτρολόγιο των 100 πλήκτρων, τότε από τη χρήση του μπορούν να εξαχθούν 100 keystroke durations και περίπου 400.000 digram latencies, εκ των οποίων οι περίπου 100.000 είναι DDDL. Αυτός είναι ένα πολύ μεγάλος αριθμός παραμέτρων που θα οδηγήσει σε συστήματα τα οποία θα απαιτούν μεγάλο χρόνο εκπαίδευσης, αλλά και εξαγωγής αποτελεσμάτων, με συνέπεια να καθίστανται ανεπαρκή στις περιπτώσεις όπου είναι αναγκαία η άμεση λήψη απόφασης. Για το λόγο αυτό ήταν επιτακτικό να χρησιμοποιηθεί μόνο ένα υποσύνολο από τις διαθέσιμες παραμέτρους. Έτσι, ύστερα από έλεγχο των αρχείων καταγραφής, εντοπίστηκαν τα πλήκτρα και τα διγράμματα τα οποία χρησιμοποιήθηκαν σχεδόν από όλους τους χρήστες και με σχετικά μεγάλη συχνότητα. Η διαδικασία αυτή κατέληξε σε 60 διάρκειες πατήματος πλήκτρου και 140 λανθάνοντες χρόνους διγράμματος.

Για την εξαγωγή των παραμέτρων σχεδιάστηκε νέο λογισμικό, με το όνομα «ISqueezeU», το οποίο δεχόταν ως είσοδο αρχεία κειμένου μορφής όπως αυτά που παραγόταν από το «IRecU» και εξήγαγε τη μέση τιμή των παραμέτρων που είχαν προεπιλεγεί, εφόσον το πλήθος εμφανίσεων του αντίστοιχου πλήκτρου ή διγράμματος υπερέβαινε την τιμή ενός κατωφλίου που επίσης είχε προκαθοριστεί. Το κατώφλι ορίστηκε στις 10 εμφανίσεις για τα πλήκτρα και στις 5 εμφανίσεις για τα διγράμματα, ενώ ο λόγος ύπαρξής του ήταν η απομάκρυνση των τιμών που δεν θα ήταν αντιπροσωπευτικές της συμπεριφοράς ενός χρήστη επί του πληκτρολογίου.

Ένα παράδειγμα της εξόδου του «ISqueezeU» παρουσιάζεται στον Πίνακα 2.

Πίνακας 2. *Παράδειγμα εξόδου του «ISqueezeU»*

Παράμ.	**Αρχείο Καταγραφής**							
	003	**113**	**127**	**155**	**198**	**209**	**218**	**227**
32	96,2	64,1	126,0	92,6	136,6	126,5	77,5	83,3
65	82,2	91,0	120,1	85,6	152,3	162,5	50,2	72,6
77	87,6	66,5	121,4	71,1	166,3	139,6	62,3	74,1
87	?	80,8	94,4	77,8	134,9	?	58,4	62,5
32-68	562,6	448,4	881,1	544,0	?	1169,8	404,0	667,8
65-73	277,1	115,5	247,2	227,2	158,0	299,3	113,6	156,0
71-82	249,9	296,3	?	303,0	148,4	211,8	284,6	180,6
77-69	356,1	145,9	222,6	236,6	190,2	356,2	198,8	145,6

Το λατινικό ερωτηματικό (?) δηλώνει ότι η συγκεκριμένη παράμετρος, στο συγκεκριμένο αρχείο καταγραφής, δεν είχε επαρκή αριθμό εμφανίσεων.

3.3 Ταξινομητής

Για την επίτευξη της καλύτερης απόδοσης του συστήματος δοκιμάστηκε ένα πλήθος ταξινομητών, συμπεριλαμβανομένων αυτών που βασίζονται σε απόσταση

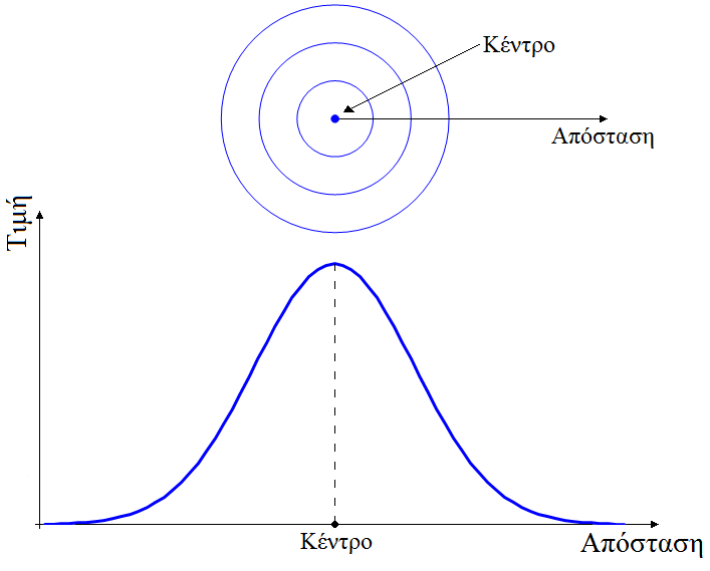
(Ευκλείδεια, Manhattan, κτλ), σε δέντρα απόφασης (decision trees), στο θεώρημα του Bayes, σε μηχανές υποστήριξης διανυσμάτων και σε νευρωνικά δίκτυα.

Η κατάληξη αυτής της διαδικασίας ήταν ένα νευρωνικό δίκτυο με συνάρτηση ακτινωτής βάσης (radial basis function network, RBFN), το οποίο επιλέχτηκε γιατί παρουσίαζε υψηλό ποσοστό πρόβλεψης και μικρό χρόνο εκπαίδευσης (training time).

Τα νευρωνικά δίκτυα με συνάρτηση ακτινωτής βάσης διατυπώθηκαν για πρώτη φορά από τους Broomhead και Lowe (1988) και μία από τις σημαντικές διαφορές που παρουσιάζουν, σε σχέση με έναν MLP, είναι ότι ως συνάρτηση μεταφοράς χρησιμοποιούν συνάρτηση ακτινωτής βάσης (radial basis function, RBF), από όπου πήραν και το όνομά τους. Η τιμή που επιστρέφει μία RBF εξαρτάται μόνο από την απόσταση της μεταβλητής από ένα σημείο, το οποίο ονομάζεται κέντρο. Όταν αυτή η απόσταση είναι μηδενική, τότε η συνάρτηση παίρνει τη μέγιστη τιμής της, ενώ όταν η απόσταση τείνει στο άπειρο, τότε η τιμή της τείνει στο μηδέν.

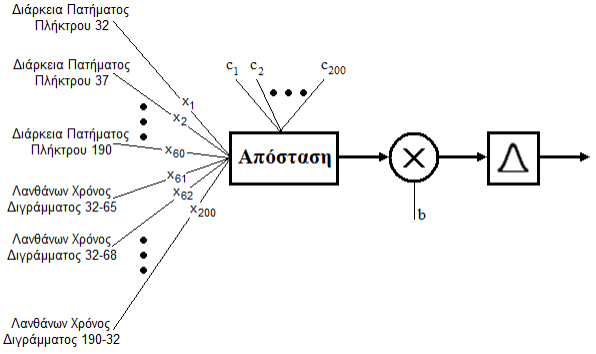
Η μορφή μιας RBF είναι παραπλήσια με αυτή του Σχήματος 1.

Σχήμα 1. *Συνάρτηση ακτινωτής βάσης*



Επίσης, η λειτουργία ενός νευρώνα RBF, η οποία αναπαριστάται στο Σχήμα 2, είναι διαφορετική από αυτή ενός perceptron.

Σχήμα 2. *Λειτουργία νευρώνα δικτύου με συνάρτηση ακτινωτής βάσης*



Οι είσοδοι του νευρωνικού δικτύου x_1, x_2, …, x_{200}, των 60 διαρκειών πατήματος πλήκτρου και των 140 λανθανόντων χρόνων διγράμματος, σχηματίζουν ένα διάνυσμα *x* 200 διαστάσεων. Το διάνυσμα αυτό εφαρμόζεται στην είσοδο του νευρώνα και υπολογίζεται η απόστασή του από ένα άλλο διάνυσμα *c*, ίδιων διαστάσεων, που είναι το διάνυσμα-κέντρο του νευρώνα. Η απόσταση μεταξύ των δύο διανυσμάτων, που συμβολίζεται ||*x*-*c*||, τυπικά λαμβάνεται από τον υπολογισμό της Ευκλείδειας απόστασης, αν και η χρησιμοποίηση της απόστασης Mahalanobis φαίνεται να αποδίδει καλύτερα.

Η υπολογισμένη απόσταση πολλαπλασιάζεται με έναν συντελεστή *b* και στο γινόμενο εφαρμόζεται η συνάρτηση ακτινωτής βάσης. Το αποτέλεσμα της συνάρτησης αποτελεί και την έξοδο του νευρώνα που γράφεται ως:

$$y_i(x) = r(\|x - c\|) \tag{1}$$

Η συνάρτηση *r* είναι ακτινωτής βάσης και επιλέγεται να δίνεται από τον τύπο:

$$r(\|x - c\|) = e^{-b\cdot\|x-c\|^2} \tag{2}$$

Ο δείκτης *i* στην εξίσωση (1) δηλώνει ότι πρόκειται για την έξοδο του *i* νευρώνα του δικτύου, αφού για την επίλυση ενός προβλήματος όπως η κατηγοριοποίηση χρηστών ανά ημερήσια χρήση υπολογιστή, με παραμέτρους τα keystroke durations και τα digram latencies, απαιτείται η χρησιμοποίηση πολλών νευρώνων που σχηματίζουν ένα νευρωνικό δίκτυο με συνάρτηση ακτινωτής βάσης.

Ο συνδυασμός των εξόδων των νευρώνων για τον υπολογισμό της τελικής εξόδου του δικτύου γίνεται με γραμμικό τρόπο, με τη βοήθεια συντελεστών *α*:

$$y(x) = \sum_{i=1}^{N} \alpha_i \cdot r(\|x - c_i\|) \tag{3}$$

Όπου *N* είναι το πλήθος των νευρώνων του δικτύου στο κρυμμένο στρώμα (hidden layer) και όπου c_i είναι το διάνυσμα-κέντρο του *i* νευρώνα.

Από τις εξισώσεις (2) και (3) προκύπτει η έξοδος του νευρωνικού δικτύου:

$$y(x) = \sum_{i=1}^{N} \alpha_i \cdot e^{-b_i \cdot \|x - c_i\|^2} \quad (4)$$

Οι συντελεστές α_i και b_i, καθώς και τα διανύσματα c_i, λαμβάνουν τέτοιες τιμές ώστε η απόδοση του ταξινομητή να βελτιστοποιηθεί.

Συγκεκριμένα, κατά τη φάση της εκπαίδευσης του συστήματος, σχηματίζονται συστάδες (clusters) των δεδομένων εισόδου. Υπενθυμίζεται ότι οι τιμές των παραμέτρων της δυναμικής της πληκτρολόγησης του κάθε δείγματος, που εφαρμόζονται ως είσοδος, ορίζουν ένα διάνυσμα 200 διαστάσεων. Σε κάθε συστάδα δεδομένων εισόδου αντιστοιχίζεται ένα διάνυσμα-κέντρο, ίδιου αριθμού διαστάσεων. Ο διαχωρισμός των δειγμάτων σε συστάδες και κατά συνέπεια ο υπολογισμός των κέντρων τους, μπορεί να γίνει με οποιονδήποτε αλγόριθμο συσταδοποίησης (clustering algorithm), όπως για παράδειγμα με αυτόν των *k* μέσων (*k*-means clustering algorithm), που χρησιμοποιήθηκε για πρώτη φορά από τον MacQueen (1967). Το πλήθος των συστάδων αποτελεί μία παράμετρο σχεδιασμού του ταξινομητή, που καθορίζει το πλήθος των νευρώνων στο κρυμμένο στρώμα του δικτύου. Τα υπολογισμένα κέντρα των συστάδων γίνονται τα κέντρα των νευρώνων.

Ο σχηματισμός των συστάδων εξαρτάται και από μία ακόμα παράμετρο σχεδιασμού, την ελάχιστη τυπική απόκλιση (minimum standard deviation) που επιτρέπεται να έχουν τα σύνολα δεδομένων που τις απαρτίζουν. Όταν η ελάχιστη τυπική απόκλιση είναι αρκετά μικρή, τότε είναι πιθανό ο ταξινομητής να δημιουργήσει συστάδες που αποτελούνται από μία μόνο είσοδο.

4. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρχικώς, ελέγχθηκε η συμπεριφορά του συστήματος για διάφορες τιμές στις παραμέτρους του ταξινομητή. Συγκεκριμένα, μετρήθηκε το ποσοστό ορθής πρόβλεψης του συστήματος για τιμές στο πλήθος των συστάδων από 10 έως και 140, και για τιμές στην ελάχιστη επιτρεπτή τυπική απόκλιση από 1,0 έως και 1,9. Για να εξασφαλιστεί ότι τα αποτελέσματα των πειραμάτων είναι ανεξάρτητα από τα δεδομένα, ακολουθήθηκε η τεχνική της διασταυρωμένης επικύρωσης *k* φορών (*k*-fold cross-validation). Πρόκειται για μία τεχνική της στατιστικής επιστήμης, η οποία εδραιώθηκε στα τέλη της δεκαετίας του 1960, όταν αρκετές δημοσιεύσεις αναφερόντουσαν σε αυτή, όπως για παράδειγμα η εργασία των Lachenbruch και Mickey (1968). Τα αποτελέσματα των πειραμάτων εμφανίζονται στον Πίνακα 3.

Τα συμπεράσματα που εξάγονται από τον Πίνακα 3 είναι, κατά πρώτον, ότι το σύστημα παρουσιάζει ένα ποσοστό επιτυχίας 68%-75% για ένα μεγάλο εύρος των τιμών των δύο παραμέτρων σχεδίασης του ταξινομητή που μελετήθηκαν. Κατά δεύτερον, ότι το σύστημα έχει την ίδια συμπεριφορά, για την ίδια τιμή ελάχιστης τυπικής απόκλισης, όταν το πλήθος των συστάδων είναι πάνω από 80. Επίσης, κάτι που δεν παρουσιάζεται σε αυτόν τον Πίνακα, είναι ότι όταν το πλήθος των συστάδων μικραίνει, αυξάνεται ο χρόνος εκπαίδευσης του ταξινομητή. Ο λόγος είναι ότι όταν ο αριθμός συστάδων είναι μεγάλος, τότε αρκετές από αυτές αποτελούνται από ένα μόνο δείγμα, με αποτέλεσμα να μην δημιουργούνται πολλοί συνδυασμοί για δοκιμή.

Αντίθετα, όταν ο αριθμός των συστάδων είναι μικρός, υπάρχουν πολλοί διαφορετικοί συνδυασμοί που μπορούν να υλοποιήσουν τη συσταδοποίηση, και άρα ο χρόνος εκπαίδευσης είναι μεγάλος. Σε κάθε περίπτωση όμως, η λειτουργία του νευρωνικού δικτύου με συνάρτηση ακτινωτής βάσης είναι ταχύτερη αυτής του MLP.

Πίνακας 3. *Ποσοστό ορθής πρόβλεψης (%) για ζεύγη τιμών αριθμού συστάδων και ελάχιστης επιτρεπτής τυπικής απόκλισης, του ταξινομητή ημερήσιας χρήσης υπολογιστή*

		Πλήθος Συστάδων							
		10	**20**	**40**	**60**	**80**	**100**	**120**	**140**
Ελάχιστη Τυπική Απόκλιση	**1,0**	71,8	69,4	68,5	71,4	70,2	70,2	70,2	70,2
	1,1	70,6	68,2	70,9	74,2	71,8	71,8	71,8	71,8
	1,2	70,2	68,6	71,8	75,0	72,2	72,2	72,2	72,2
	1,3	69,0	70,2	73,0	74,6	72,2	72,2	72,2	72,2
	1,4	68,9	68,5	72,6	73,8	71,4	71,4	71,4	71,4
	1,5	68,6	70,1	73,0	73,4	70,2	70,2	70,2	70,2
	1,6	68,5	68,9	73,0	72,6	69,4	69,4	69,4	69,4
	1,7	66,5	68,9	71,8	72,2	67,7	67,7	67,7	67,7
	1,8	66,9	69,4	72,2	70,6	66,5	66,5	66,5	66,5
	1,9	66,1	70,2	72,6	69,0	64,1	64,1	64,1	64,1

Μια πιο εστιασμένη εικόνα των αποτελεσμάτων παρουσιάζεται στον Πίνακα 4, με τιμές παραμέτρων του ταξινομητή 100 συστάδες και 1,3 ελάχιστη επιτρεπτή τυπική απόκλιση.

Πίνακας 4. *Αποτελέσματα πρόβλεψης ημερήσιας χρήσης υπολογιστή*

	Προβλέφθηκαν ως					
Ημερήσια Χρήση Υπολογιστή	**0-1**	**1-2**	**2-4**	**4-6**	**6+**	**Ποσοστό Επιτυχίας ανά Κλάση**
0-1	16	2	0	1	4	69,6%
1-2	2	30	9	0	5	65,2%
2-4	3	1	39	3	8	72,2%
4-6	2	0	5	31	0	81,6%
6+	3	2	16	3	63	72,4%
Ποσοστό Επιτυχίας ανά Πρόβλεψη	61,5%	85,7%	56,5%	81,6%	78,8%	

Όπως φαίνεται, από το σύνολο των 248 αρχείων, τα 179 προβλέφθηκαν ορθώς, κάτι που σημαίνει ένα συνολικό ποσοστό επιτυχίας που υπερβαίνει το 72%, το οποίο είναι κατά πολύ υψηλότερο από το 20% της τυχαίας πρόβλεψης. Τα υπόλοιπα στατιστικά μεγέθη που συνοδεύουν το συγκεκριμένο πείραμα είναι ο σταθμισμένος μέσος όρος της ακρίβειας (precision) στα 0,74, της ανάκλησης (recall) στα 0,72 και της μέτρησης F (F-measure) στα 0,73. Σε ότι αφορά την περιοχή κάτω από την καμπύλη ROC (ROC Area), ο σταθμισμένος μέρος όρος της ανέρχεται σε 0,83. Τέλος, η τιμή του μέσου απόλυτου σφάλματος (mean absolute error) είναι 0,1113 και ο συντελεστής κάπα του Cohen είναι 0,6372.

Στα συμπεράσματα αυτής της έρευνας συμπεριλαμβάνεται και η ευθυγράμμιση των προδιαγραφών λειτουργίας του ταξινομητή με την πειραματική συμπεριφορά του. Δηλαδή, όπως αναμενόταν, το νευρωνικό δίκτυο με συνάρτηση ακτινωτής βάσης κατηγοριοποίησε τους χρήστες σύμφωνα με την ημερήσια χρήση υπολογιστή με υψηλό ποσοστό επιτυχίας και σε σύντομο χρονικό διάστημα.

5. ΣΥΝΟΨΗ – ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία η δυναμική της πληκτρολόγησης, δηλαδή η λεπτομερής χρονική καταγραφή των ενεργειών ενός χρήστη επί του πληκτρολογίου QWERTY, χρησιμοποιήθηκε για την κατηγοριοποίηση χρηστών. Με την κατηγοριοποίηση χρήστη, τη διαδικασία κατά την οποία ένας χρήστης υπολογιστή εντάσσεται σε μία κλάση, επιτυγχάνεται η πρόβλεψη κάποιων εγγενών ή επίκτητων χαρακτηριστικών του, όπως το φύλο, η ηλικία, η προτίμηση χεριού, το μορφωτικό επίπεδό του, κ.α. Για τον εντοπισμό των χαρακτηριστικών των χρηστών έχουν προταθεί και άλλες μέθοδοι, όπως η εξέταση των φωτογραφιών, των κειμένων και του κοινωνικού δικτύου τους, κάθε μία όμως από αυτές αντιμετωπίζει κάποιους περιορισμούς, με αποτέλεσμα την αδυναμία γενίκευσής της. Αντίθετα, με τη δυναμική της πληκτρολόγησης απορρέει μέθοδος που είναι ανεξάρτητη ομιλούμενης γλώσσας, δεν επεξεργάζεται ευαίσθητα ή και προσωπικά δεδομένα και απαιτεί μόνο τη παραγωγή του πιο συνηθισμένου μέσου επικοινωνίας μεταξύ των χρηστών, του κειμένου.

Το χαρακτηριστικό που εξετάστηκε ήταν η ημερήσια χρήση υπολογιστή, από το οποίο δεν προκύπτουν άμεσα συμπεράσματα για τα χαρακτηριστικά του χρήστη, αλλά χρήσιμες ενδείξεις για το φύλο, την ηλικία και το ετήσιο εισόδημά του, ενώ ταυτόχρονα παρέχονται και πληροφορίες εγκληματολογικού ενδιαφέροντος, αφού για παράδειγμα, η εγκληματολογική έρευνα θα απέκλειε όλους εκείνους τους υπόπτους που το πλήθος των ωρών ενασχόλησής τους με υπολογιστή ανά ημέρα, θα τους κατέτασσε στους μη εξειδικευμένους χρήστες και άρα όχι ικανούς για μια κακόβουλη ηλεκτρονική επίθεση. Η απόκτηση τέτοιου είδους πληροφοριών, σε συνδυασμό με άλλες που προκύπτουν και αυτές από την κατηγοριοποίηση χρηστών διά μέσου δυναμικής της πληκτρολόγησης, όπως για παράδειγμα το φύλο, η ηλικία, το μορφωτικό επίπεδο, το επικρατές χέρι και άλλα χαρακτηριστικά του ατόμου που διέπραξε μια κακόβουλη ενέργεια, θα καθιστούσαν την προτεινόμενη μέθοδο πολύτιμο εργαλείο στα χέρια των ειδικών της ψηφιακής εγκληματολογίας.

Το σύστημα που παρουσιάζεται εκμεταλλεύεται 60 διάρκειες πατήματος πλήκτρου και 140 λανθάνοντες χρόνους διγράμματος, ενώ υλοποιείται με τη βοήθεια ενός νευρωνικού δικτύου με συνάρτηση ακτινωτής βάσης. Τα αποτελέσματα έδειξαν ένα ποσοστό ορθής πρόβλεψης που ανέρχεται έως και το 75%, που είναι εμφανώς υψηλότερο από το 20% της τυχαίας πρόβλεψης.

Η πρωτοτυπία της εργασίας έγκειται στην εκμετάλλευση των παραμέτρων της δυναμικής της πληκτρολόγησης στην κατηγοριοποίηση χρηστών, καθώς επίσης και στην πρώτη απόπειρα εύρεσης της ημερήσιας χρήσης υπολογιστή. Σημαντικό συμπέρασμα αποτελεί επίσης το ότι οι διάρκειες πατήματος πλήκτρου και οι λανθάνοντες χρόνοι διγράμματος μπορούν να χρησιμοποιηθούν για την εύρεση κάποιων χαρακτηριστικών ενός άγνωστου χρήστη.

ABSTRACT

Internet is a tool where billions of people are now connected for purposes of communication, entertainment, work and education. But alongside the benefits, many dangers are lurking that may harm the users financially, morally and socially. Many of these risks come from malicious users and one of their “weapons” is that they can retain their anonymity. In this paper we propose a method that exploits the data generated by the way people type. The reason is to find some characteristics of users, so that to remove the complete anonymity and to be able to alert the unsuspecting users for the probability that their interlocutors misrepresent their characteristics. The proposed system was implemented with artificial neural networks and correct prediction exceeded 75%, well above than 20% of random prediction.

ΑΝΑΦΟΡΕΣ

Alrashed, H. F., & Berbar, M. A. (2013). Facial gender recognition using eyes images. *International Journal of Advanced Research in Computer and Communication Engineering, 2*(6), 2441-2445.

Bartlow, N., & Cukic, B. (2006). Evaluating the reliability of credential hardening through keystroke dynamics. *In Proceedings of 17th International Symposium on Software Reliability Engineering*. doi:10.1109/issre.2006.25.

Broomhead, D., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems, 2*, 321–355.

Cheng, N., Chandramouli, R., & Subbalakshmi, K. (2011). Author gender identification from text. *Digital Investigation, 8*(1), 78-88. doi:10.1016/j.diin.2011.04.002.

Choobeh, A. K. (2012). Improving automatic age estimation algorithms using an efficient ensemble technique. *International Journal of Machine Learning and Computing 2*(2), 118-122. doi:10.7763/ijmlc.2012.v2.99.

Daily computer usage in Great Britain by age 2006-2015. (n.d.). Retrieved from http://www.statista.com/statistics/275996/daily-computer-usage-penetration--in-great-britain-by-age/ (accessed on 15/04/2016).

Damayanti, F., & Rachmad, A. (2016). Recognizing gender through facial image using support vector machine. *Journal of Theoretical and Applied Information Technology, 88*(3), 607-612.

Douhou, S., & Magnus, J. R. (2009). The reliability of user authentication through keystroke dynamics. *Statistica Neerlandica, 63*(4), 432-449. doi:10.1111/j.1467-9574.2009.00434.x.

Hassan, S., Selim, M., & Zayed, H. (2013). User authentication with adaptive keystroke dynamics. *International Journal of Computer Science Issues, 10*(4), 127-134.

Hewahi, N., Olwan, A., Tubeel, N., EL-Asar, S., & Abu-Sultan, Z. (2010). Age estimation based on neural networks using face features. *Journal of Emerging Trends in Computing and Information Sciences, 1*(2), pp. 61-67.

Hofferth, S., Flood, S., & Sobek, M. (2013). American time use survey data extract system: Version 2.4 [Machine-readable database]. Maryland Population Research Center, University of Maryland, College Park, Maryland, and Minnesota Population Center, University of Minnesota, Minneapolis, Minnesota.

Hosseinzadeh, D., & Krishnan, S. (2008). Gaussian mixture modeling of keystroke patterns for biometric applications. *IEEE Transactions on Systems, Man, and Cybernetics, 38*(6), 816-826. doi:10.1109/tsmcc.2008. 2001696.

Internet live stats. (n.d.). Retrieved from http://www.internetlivestats.com (accessed on 02/02/2017).

Kumar, A., Patwari, A., & Sabale, S. (2014). User authentication by typing pattern for computer and computer based devices. *International Journal of Advanced Research in Computer and Communication Engineering, 3*(10), 8132-8134. doi:10.17148/ijarcce.2014.31011.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10*(1), 1. doi:10.2307/1266219.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297.

Mukherjee, A., & Liu, B. (2010). Improving gender classification of blog authors. *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 207-217.

Patil, R., & Renke, A. (2016). Keystroke dynamics for user authentication and identification by using typing rhythm. *International Journal of Computer Applications, 144*(9), 27-33. doi:10.5120/ijca2016910432.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. *In Proceedings of 2nd International Workshop on Search and Mining User-Generated Contents*, 37-44. doi:10.1145/1871985.1871993.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. *In Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 199-205.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321-326. doi:10.1089/1094931041291295.

Wankhede, S., & Verma, S. (2014). Keystroke dynamics authentication system using neural network. *International Journal of Innovative Research and Development, 3*(1), 157-164.

Winn, J., & Heeter, C. (2009). Gaming, gender, and time: Who makes time to play? *Sex Roles, 61*(1-2), 1-13. doi:10.1007/s11199-009-9595-7.

Yi, D., Lei, Z., & Li, S. Z. (2014). Age estimation by multi-scale convolutional network. *In Proceedings of 12th Asian Conference of Computer Vision,* 144-158. doi:10.1007/978-3-319-16811-1_10.

Zhao, Y. (2006). Learning user keystroke patterns for authentication. *In Proceedings of World Academy of Science, Engineering and Technology, 14*, 65-70.

εργασίες

στα αγγλικά

Greek Statistical Institute
Proceedings of the 30th Panhellenic Statistics Conference (2017), pp. 97−108

BAYESIAN TESTING FOR ASSOCIATION MODELS IN CONTINGENCY TABLES USING POWER PRIOS

A. Mantzouni
Department of Statistics, Athens University of Economics and Business, Greece
aikmantz@aueb.gr

ABSTRACT

In this paper we illustrate a comprehensive Bayesian analysis of association models, involving suitable choices of prior parameters, estimation, model determination, as well as the allied computational issues for contingency tables. More specifically, the use of the conjugate prior distribution in Bayesian analysis sometimes is problematic due to the well known sensitivity of the posterior model odds and the Barlett-Lindley paradox (Lindley, 1957, Barlett, 1957). This fact leads to the utilization of power prior approach. We advocate sensible values for the prior parameter on the full table and the corresponding induced values for the rest of the association models. We produce imaginary set of data and a pre-prior with all parameter equal. The unnormalized prior distribution can be obtained by the product of the likelihood raised to a power, multiplied by the pre-prior distribution. Unit information interpretation priors are used as a yardstick in order to identify and interpret the effect of any other prior distribution used. The posterior distributions of the association models parameters based on power prior setup, are obtained using simple Markov chain Monte Carlo (MCMC) schemes. Evaluation of the models under consideration and related Bayesian tests can be obtained using MCMC based on marginal likelihood estimations (see for example Perrakis *et al.*, 2014), Laplace approximations and Laplace Metropolis estimators. For this class of models, we present real data sets to demonstrate the proposed methodology.

Keywords: Contingency Tables, Association Models, Bayes Factor, Power Prior, Laplace Approximation, Laplace Metropolis Estimator.

1. INTRODUCTION

Let X and Y be two categorical variables of $I \geq 2$ and $J \geq 2$ levels, respectively, that are cross-classified in a $I \times J$ contingency table and $\boldsymbol{n} = (n_{ij})$, with n_{ij} being the observed frequency for cell (i, j), $i = 1, \ldots, I$, $j = 1, \ldots, J$. In the following we assume independent Poisson sampling distribution for each cell, $n_{ij} \sim Poisson(\lambda_{ij})$. The total sample size of this table is denoted by

$N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. In two way contingency tables analysis, a popular class of models for describing the structure of the association between the two categorical variables are the association models. This class of models assign scores to the classification variables, which can be either fixed and prespecified or unknown (to be estimated) parameters.

The association models (Goodman, 1985) can be expressed as

$$log(\lambda_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Where λ_0 is the overall measure of average log-counts, λ_i^X and λ_j^Y are the marginal effect terms for factors X (rows) and Y (columns). The product of row scores $\mu = \{\mu_1, .., \mu_I\}$ and column scores $\nu = \{\nu_1, .., \nu_I\}$ multiplied by parameter ϕ, replaces the interaction term of the saturated log-linear model, $\lambda_{ij}^{XY} = \phi\mu_i\nu_j$. The parameter ϕ is redundant and is a global measure of association under certain parametrization.

The types of association models, depending on the type of the row and column parameter scores are:

- The Linear by Linear (LL) model, with fixed row and column scores. Uniform (U) is the most characteristic one with equidistant row and column scores;
- Row effect association model (R), with unknown row and fixed column scores;
- Column effect association model (C), with unknown column and fixed row scores;
- Row-Column association model (RC), with both row and column scores to be parameters under estimation.

For identifiability purposes, the sum-to-zero constraints are impose on row and column main effects

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0$$

Iliopoulos *et al.* (2009) adopted alternative parametrizations for the RC model

$$\mu_1 = \nu_1 = 0 \quad \text{and} \quad \mu_I = \nu_J = 1,$$

which fix the minimum and maximum score for each classification variable. By this parametrization, the redundant score parameters are set to be fixed and as a consequence the derivation of posterior distribution for them is avoided.

Bayesian independence tests in $I \times J$ contingency tables are based on posterior the model odds and the Bayes factor. Bayes factor provides a way to formally compare two competing models. By Bayes theorem, the posterior odds in favor

of M_0 versus M_1 is given by

$$PO_{01} = \frac{f(M_0 \mid \boldsymbol{n})}{f(M_1 \mid \boldsymbol{n})} = \frac{f(\boldsymbol{n} \mid M_0)}{f(\boldsymbol{n} \mid M_1)} \times \frac{f(M_0)}{f(M_1)} = B_{01} \times \frac{f(M_0)}{f(M_1)}.$$

The Bayes factor of model M_0 versus model M_1 is defined as the ratio of Bayesian marginal likelihoods of model M_0 versus model M_1 (Jeffreys, 1961):

$$B_{01} = \frac{f(\boldsymbol{n}|M_0)}{f(\boldsymbol{n}|M_1)}.$$

The Bayes factor quantifies the evidence for and against the hypothesis of independence in contingency tables, when the prior model probabilities $f(\boldsymbol{\mu})$ of models $\boldsymbol{\mu} \in \{M_0, M_1\}$ are equal, i.e. $f(M_0) = f(M_1)$.

Posterior model probabilities and Bayes factor are highly sensitive to the prior specification of the model parameters. This result was reported after the publication of Lindley (1957), who reported a surprising behavior of the Bayes factor. When the sample size N increases, then the Bayes factor also increases and tends to infinity, fully supporting the simpler hypothesis:

$$N \to \infty \quad \Rightarrow \quad B_{01} \to \infty.$$

This behavior is known as the Lindley's paradox. Subsequently, motivated by the work of Lindley, Barlett (1957) extended this paradox by observing that the prior variance of the additional parameters in nested models comparisons (when $M_0 \subseteq M_1$), also massively affects the Bayes factor B_{01} as it tends to infinity:

$$Var(\lambda_{ij}) \to \infty \quad \Rightarrow \quad B_{01} \to \infty.$$

This behavior is known by the name of Jeffreys or Lindley's or Barlett paradox. We use power priors to avoid the effect of the Lindey's paradox. We account imaginary data $\boldsymbol{n^*} = (n^*_{ij})$ and we weight them to account for one data point as a reasonable "low-information" choice.

2. METHODS OF MARGINAL LIKELIHOOD COMPUTATION

Bayesian model comparison via the Bayes factor, posterior model probabilities and odds (Kass and Raftery, 1995), requires the computation of the Bayesian marginal likelihood given by

$$f(\boldsymbol{n}|M) = \int f(\boldsymbol{n}|\boldsymbol{\vartheta}, M) f(\boldsymbol{\vartheta}|M) d\boldsymbol{\vartheta}, \tag{1}$$

where $f(\boldsymbol{\vartheta}|M)$ is the density of the model specific parameter vector $\boldsymbol{\vartheta} \in \Omega \subset \mathbb{R}^p$. Historically, the integration required for calculating the equation (1) has been done by taking advantage of conjugacy by assuming approximate posterior normality

or by using numerical quadrature, the Laplace method or Monte Carlo integration (see Kass and Raftery, 1995). Recently, it has become possible to estimate a wider range of models, using posterior simulation methods such as Monte Carlo sampling methods to avoid the analytical computation of the marginal likelihood (Neal, 2000, Perrakis *et al.*, 2014). Lartillot and Philippe (2006) introduced a technique called thermodynamic integration to approximate the marginal likelihood. A similar method, stepping-stone-sampling (Xie *et al.*, 2011, Fan *et al.*, 2011), has more recently been proposed (see also Baele *et al.*, 2012, Baele and Lemey, 2013, Friel *et al.*, 2014, for a summary and comparison of these methods). The previous ways of calculating integrated likelihoods often cannot be used for models estimated via MCMC or their posterior simulation methods. There is a variety of methods to compute the marginal likelihood, but simplicity is not the strongest point of most methods. In this paper two methods for estimating the marginal likelihood are presented, the Laplace approximation approach and the utilization of the Laplace-Metropolis estimator.

2.1 Laplace Approximation

The Laplace approximation for an integral of the form $\int e^{h(u)}du$ is found using a Taylor series expansion of a real-valued of a P-dimensional vector u. Rosenkranz (1992) found that the Laplace method produces much more accurate estimates of the marginal likelihood than posterior simulation for a variety of models. The marginal likelihood for contingency tables can be approximated by

$$\log f\left(\mathbf{n}|M\right) \approx \frac{d_m}{2}\log(2\pi) + \frac{1}{2}\log \mid H^* \mid + \log f\left(\boldsymbol{\vartheta}_M^* \mid M\right) + \log f\left(\mathbf{n} \mid \boldsymbol{\vartheta}_M^*, M\right),$$

where $\boldsymbol{\vartheta}_M^*$ is the posterior mode of model M, and H^* is minus the inverse Hessian matrix of the log-posterior density evaluated at the posterior mode.

2.2 Laplace-Metropolis Estimator

Lewis and Raftery (1997) describe a way to use posterior simulation output to estimate integrated likelihoods. Laplace method is often not applicable because the derivatives that it requires are not easily available. This is particularly true for complex models of the kind for which posterior simulation, especially MCMC, is often used. The idea of the Laplace-Metropolis estimator is to get around the limitations of the Laplace method by using posterior simulation to estimate the quantities it needs. To avoid analytic calculation of H^* and $\boldsymbol{\vartheta}_M^*$ Laplace-Metropolis estimator proposed $\boldsymbol{\vartheta}_M^*$ to be estimated by the posterior mean $\bar{\boldsymbol{\vartheta}}_M$ and H^* by the posterior variance-covariance matrix of $\boldsymbol{\vartheta}$ obtained by the MCMC output.

$$\log f\left(\mathbf{n}|M\right) \approx \hat{f}(\mathbf{n}|M),$$

$$\hat{f}(\mathbf{n}|M) = \frac{d_m}{2}\log(2\pi) + \frac{1}{2}\log \mid \boldsymbol{S}_M \mid + \log f\left(\bar{\boldsymbol{\vartheta}}_M \mid M\right) + \log f\left(\mathbf{n} \mid \bar{\boldsymbol{\vartheta}}_M, M\right)$$

There are several possible ways of estimating $\boldsymbol{\vartheta}^*$ from an MCMC sample:

- Estimate $\boldsymbol{\vartheta}^*$ as that $\boldsymbol{\vartheta}$ in the sample at which $h(\boldsymbol{\vartheta}) = f(\boldsymbol{n}|\boldsymbol{\vartheta}, M)f(\boldsymbol{\vartheta}|M)$ achieves its maximum.
- Estimate the components of $\boldsymbol{\vartheta}^*$ by finding the componentwise posterior means.
- Estimate the components of $\boldsymbol{\vartheta}^*$ by finding the componentwise posterior medians.
- Estimate $\boldsymbol{\vartheta}^*$ by finding the multivariate median.

The first of these methods is the simplest conceptually and usually the most accurate. However, it involves calculating the likelihood many times and therefore it might be computationally expensive. In such cases, the multivariate median might be used instead, which not require too much computing resources. Moreover, the MCMC estimated posterior median is more robust than the corresponding posterior mean, which is influenced by outliers. Furthermore the median is more accurate proxy of the model than the mean for a wide range of distributions (Johnson and Kotz, 1985). The other quantity required for the calculation of Laplace-Metropolis estimator is H^*. This is asymptotically equal to the posterior variance matrix, and we could estimate it by the sample covariance matrix of the posterior simulation output. However, because MCMC trajectories take occasional distant excursions, it is recommended to use a robust estimator of the posterior variance matrix.

2.3 Comparison of the two methods

Laplace approximation is obtained by using any optimization method, while Laplace-Metropolis estimator is based on the output of a MCMC algorithm. The first method is faster and without any Monte Carlo error, but it is not always applicable due to computational problems. The second approach is more computationally demanding, but the implementation is direct when the MCMC sample is available. Both estimators are approximations of the Bayesian marginal likelihood and therefore they require large enough sample size in order to obtain accurate results. Therefore they will not be accurate for small or sparse datasets. Both approaches have some regularities conditions (see Kass and Wasserman, 1996). These approximations are more efficient when the posterior distributions are symmetric, otherwise transformations should be applied (lg, logit or Box-Cox).

3. POWER PRIOS

Prior elicitation is one of the most important issues in Bayesian data analysis. When no prior information is available, a non-informative prior such as uniform prior or Jeffreys prior can be used (see Kass and Wasserman, 1996) for a list of such non-informative priors. However, real prior information such as historical data or data from previous similar studies are often available in applied research, where the investigator has access to previous studies measuring the same response and covariates as the current study. In experiments conducted over time, data from previous time periods can often be used as prior information. It is natural to incorporate the historical data into the current study by quantifying it with a suitable prior distribution on the mode parameters. Power prior distributions are based on the idea of raising the likelihood function of historical data to a power w, where $0 \leq w \leq 1$. Power prior approach of Ibrahim and Chen (2000) and Chen (2000) is adopted to advocate sensible values for the prior parameters on contingency table. The initial idea of the power prior can be traced to Diaconis and Ylvisaker (1979) and Morris (1983), where they studied conjugate priors for exponential families. However, these two authors considered only the situation with the power w be a fixed constant.

3.1 Methodology

We adopt the idea based on the power prior approach of Ibrahim and Chen (2000) and Chen *et al.* (2000) in order to advocate sensible values for the prior parameters on the full table and the corresponding induced values for the rest of the association models. Let us consider imaginary set of data represented by the frequency table $\boldsymbol{n^*} = (n^*_{ij})$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$ of total sample size $N^* = \sum_{i \in I} \sum_{j \in J} n^*_{ij}$ and a normal ″pre-prior‟with all parameters equal. Then the unnormalized prior distribution can be obtained by the product of the likelihood of $\boldsymbol{n^*}$ raised to a power w multiplied by the "pre-prior" distribution.

$$f(\boldsymbol{\vartheta}) \propto f(\boldsymbol{n^*}|\boldsymbol{\vartheta})^w \times f(\boldsymbol{\vartheta^*})$$

The parameter w is used to specify the steepness of the prior distribution and the weight of belief on each prior observation

- $w = 1$ each imaginary observation has the same weight as the actual observations
- $w < 1$ give less weight to each imaginary observation
- $w > 1$ will increase the weight of believe on the prior/imaginary data
- $w = 1/N^*$ the prior data $\boldsymbol{n^*}$ will account information of one data point and when we use Uniform improper pre-prior then we have 1 unit of prior information towards the simple model of uniformity across all cells.
- $w = 1$, $N^* = N$ and the pre-prior is improper uniform then both the prior and data will account for 50% of the information used in the posterior.

For $w = 1/N^*$ the prior data $\boldsymbol{n^*}$ will account for information of one data point. This prior set-up will called unit information prior (UIP). When no information is available, the choice of equal cell frequencies $(n^*_{ij}) = \boldsymbol{n^*}$ for the imaginary data in order to support the simplest possible model under consideration.

The proposed methodology for association models in contingency tables has five steps.

- **Step 1:** no information is available, set equal cell frequencies $(n^*_{ij}) = \boldsymbol{n^*} = \frac{N}{IJ}$, $N^* = n^* \times IJ = N$ and $w = \dfrac{1}{N^*} = \dfrac{1}{n^* \times IJ}$
- **Step 2:** set a normal pre-prior with large variance and multiply with the Poisson likelihood for imaginary data $\boldsymbol{n^*}$
- **Step 3:** compute posterior mean and posterior variance from the MCMC output for each model
- **Step 4:** For the actual data, use a normal prior for each parameter with mean and variance obtained by Step 2. The variance is multiplied by $\frac{1}{w}$ (N for the Unit information). This is an approximation of the desired power prior which considerably simplifies all computations. Steps 2-4 can be avoided by analytic computation of the power prior which is currently work in progress
- **Step 5:** compute the marginal distribution with Laplace approximation and Laplace-Metropolis estimator and compute the Bayes factor.

4. ILLUSTRATED EXAMPLE

The classical dataset of Maxwell (1961), in which the severity of dreams disturbance of 223 boys is crossclassified with their age, has been used to illustrate the proposed methodology. Maxwell discusses an analysis of a 5×4 contingency table giving the number of boys with four different rating for disturbed dreams in five different age of groups, see Table 1. The higher the rating the more the boy suffers from disturbed dreams.

Age group	Disturbance (from low to high)				Total
	1	2	3	4	
5-7	7	4	3	7	21
8-9	10	15	11	13	49
10-11	23	9	11	7	50
12-13	28	9	12	10	59
14-15	32	5	4	3	44
Total	100	42	41	40	223

Table 1: *Cross-classification of 223 boys by severity of disturbances of dreams and age.*

We set all cell of imaginary data equal to one and impose a normal non-informative pre-prior with large variance. The posterior mean and posterior vari-

ance for all parameters in each model using the imaginary data are estimated via MCMC. These values are used to build an approximation of the power prior. The results of the new MCMC output are listed in Table 2 along with the estimated log-marginal likelihood with the Laplace approximation and Laplace-Metropolis estimator approach. The two approaches are very close supporting the same model.

Mj	**Model**	**log-marginal**	
		Laplace	**Laplace-Metropolis**
1	Independence (I)	−91.399	−91.296
2	Uniform(U)	−90.167	−90.596
3	Row (R)	−103.771	−103.159
4	Column (C)	−97.652	−97.096
5	Row-Column (RC)	−107.365	−107.446
6	Saturated (S)	−131.665	−131.253

Table 2: *Estimated the logarithm of marginal likelihood for all the competitive models with the two approaches.*

After implementing the proposed methodology using $\mu_{min} = \nu_{min} = 0$, $\nu_{max} = 1$ and $\phi = 1$ parametrization for the RC model, only two models were found with posterior model probabilities higher than 1%. Results for these models are summarized in Table 3. All model comparison measures indicate that the Uniform model is the best. According to Table 3 the highest probability model $(77,4\%)$ with Laplace Metropolis estimator approach and $(66,8\%)$ with Laplace Approximation approach, is the Uniform association model with fix row and column scores. The independence model is supported as the second best one but with considerably lower probabilities (0.331 and 0.226 for Laplace approximation and Laplace-Metropolis approach, respectively).

Mj	**Model**	**Log-BF_{j2}**		**Posterior Probabilities**	
		Laplace	***Laplace-Metropolis***	***Laplace***	***Laplace-Metropolis***
1	Independence (I)	−0.7	−1.2	0.331	0.226
2	Uniform (U)	0.0	0.0	0.668	0.774
3	Row (R)	−12.6	−13.6	<0.01	<0.01
4	Column (C)	−6.5	−7.5	<0.01	<0.01
5	Row-Column (RC)	−16.8	−17.2	<0.01	<0.01
6	Saturated (S)	−40.7	−41.5	<0.01	<0.01

Table 3: *Estimated the logarithm of Bayes factor and the posterior model probabilities for all the competitive models with the two approaches.*

All models with posterior probability lower than 1% (provided in Table 3), when compared with the model of the highest probability, provide "not worth than a bare mention" evidence in favor of the latter, according to Kass and Raftery

(1995) evaluation table for Bayes factor.

5. CONCLUSION

Both Laplace approximation and Laplace-Metropolis estimator provide a reasonable solution for low information estimation of marginal likelihood. The utilization of power priors provides good argument for a reasonable prior and avoids the effect of Lindley's paradox. Moreover,we achieve compatibility of priors across models under consideration due to the use of common imaginary data across models. In the future we will focus on the implementation of other alternative estimation methods for the computation of the marginal likelihood, for example by using the Monte Carlo estimate proposed by Perrakis *et al.* (2014) or the Chib's (1995) marginal likelihood estimator. Our goal is to extend the methodology to alternative scenarios of imaginary data and apply extended simulation study to research the behavior of the methods. Explore the analytic computation of the power prior for the steps 2-4 of the algorithm is in process.

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή παρουσιάζεται μια ολοκληρωμένη Μπεϋζιανή ανάλυση των μοντέλων συνάφειας, η οποία περιλαμβάνει μια πρόταση για την κατάλληλη επιλογή των παραμέτρων της εκ των προτέρων κατανομής, εκτίμηση και καθορισμός των μοντέλων, καθώς και επίλυση βασικών υπολογιστικών ζητημάτων. Πιο συγκεκριμένα, η χρήση συζυγών εκ των προτέρων κατανομών στη Μπεϋζιανή ανάλυση μερικές φορές μπορεί να αποβεί προβληματική, λόγω της ευαισθησίας των εκ των υστέρων πιθανονήτων των μοντέλων (posterior odds) και την επίδραση του παραδόξου του Lindley-Barlett (Lindley, 1957, Barlett, 1957). Αυτό το γεγονός οδήγησε στην χρήση μιας νέας προσέγγισης, αυτής των εκ των προτέρων κατανομών δύναμης (power priors). Για το σκοπό αυτό παράγουμε ένα πλασματικό σετ δεδομένων και ορίζουμε μια εκ των προτέρων κατανομή για τα πλασματικά αυτά δεδομένα (pre prior) με όλες τις παραμέτρους ίσες. Η μη κανονικοποιημένη εκ των προτέρων κατανομή προκύπτει από το γινόμενο της πιθανοφάνειας υψωμένο σε μια δύναμη, πολλαπλασιασμένο με την εκ των προτέρων κατανομή των πλασματικών δεδομένων. Εκ των προτέρων κατανομές μοναδιαίας ερμηνευτικής πληροφορίας χρησιμοποιούνται ως μέτρο σύγκρισης με σκοπό να καθοριστεί και να ερμηνευθεί η επίδραση οποιασδήποτε άλλης εκ των προτέρων κατανομής που χρησιμοποιείται. Η εκ των υστέρων κατανομή των παραμέτρων των μοντέλων συνάφειας με την προσέγγιση των εκ των προτέρων κατανομών δύναμης (power priors), προκύπτει με την εφαρμογή MCMC μεθόδων. Η αξιολόγηση των εξεταζόμενων μοντέλων και οι αντίστοιχοι Μπεϋζιανοί έλεγχοι υποθέσεων πραγματοποιούνται με τη χρήση MCMC μεθόδων, οι οποίες βασίζονται στην εκτίμηση της περιθώριας κατανομής (βλ. για παράδειγμα Perrakis *et al.*), στην προσεγγιστική μέθοδο του Laplace καθώς και στον εκτιμητή Laplace-Metropolis.

Γι΄ αυτήν την κλάση μοντέλων θα παρουσιάσουμε πραγματικά σετ δεδομένων για να δείξουμε τον τρόπο εφαρμογής αλλά και την καλή λειτουργία της προτεινόμενης μεθοδολογίας.

REFERENCES

Agresti, A. (2013). *Caterorical Data Analysis*, 3rd edition, Wiley & Sons.

Agresti A., Chuang C. and Kezouh A. (1987). Order-restricted score parameters in association models for contingency tables. *Journal of the American Statistical Association* **82**, 619-623.

Agresti A. and Chuang C. (1989). Model-based Bayesian methods for testing call proportions in cross-classification tables having ordered categories. *Computational Statistics & Data Analysis* **7**, 245-258.

Baele, G. and Lemey, P. (2013). Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. *Bioinformatics* **29**, 1970?1979.

Baele, G., Lemey, P., and Vansteelandt, S. (2013). Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. *BMC bioinformatics* **14**, 85.

Bartlett, M. (1957). Comment on D.V. Lindley's Statistical Paradox. *Biometrika*, **44**, 533-534.

Chen, M.H., Ibrahim, J.G. and Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society. Series B* **61**, 223-242.

Chen, M.H., Ibrahim, J.G. and Shao, Q.M. (2000). Power Prior Distributions for Generalized Linear Models. *Journal of Statistical Planning and Inference* **84**, 121?137.

Chib, S. (1995), Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association* **90**, 1313?1321.

Chuang, C. (1982). Empirical Bayes methods for a two-way multiplicative-interaction model. *Communications in Statistics: Theory and Methods* **11**, 2977–2989.

Diaconis, P. and Ylvisaker, D. (1979). Conjucate priors for exponential families. *Annals of Statistics* **7**, 269-281.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2011). Choosing among partition models in bayesian phylogenetics. *Molecular Biology and Evolution* **28**, 523?532.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 589?607.

Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing* **24**, 709? 723.

Ibrahim, J.G. and Chen, M.H. (2000). Power Prior Distributions for Regression Models. *Statistical Science* **15**, 46-60.

Jeffreys, H. (1961). Theory of Probability, *3rd edition Oxford Classic Texts in the Physical Sciences* Oxford University Press, Oxford.

Johnson, N. L., and Kotz, S. (1985).*Encyclopedia of Statistical Sciences* (Vol. 5), New York: Wiley.

Iliopoulos, G., Kateri, M. and Ntzoufras, I. (2007). Bayesian estimation of unrestricted and order-restricted association model for a two-way contingency table. *Computational Statistics and Data Analysis* **51**, 4643-4655.

Iliopoulos, G., Kateri, M. and Ntzoufras, I. (2009). Bayesian model comparison for the order restricted RC association model. *Psychometrika* **74**, 561-587.

Kass, R.E. and Raftery A. (1995). Bayes Factors. *Journal Journal of the American Statistical Association* **90**, 773-795.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343-1370.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R,* Birkhäuser.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology* **55**, 195.

Lewis S. and Raftery A. (1997). Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator. *Journal of the American Statistical Association* **92**, 648-655.

Lindley, D.(1957). A Statistical Paradox. *Biometrika*, **44**, 187-192.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Associatio*, **78**, 47?65.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models.*Journal of computational and graphical statistics* **9**, 249?265

Ntzoufras, I. (2009.) *Bayesian Modeling Using WinBUGS,* 1st ed. New York: John Wiley & Sons.

Ntzoufras, I. and Tarantola, C. (2012). Bayesian analysis of graphical models of marginal independence for three way contingency tables. *Technical report, Department of Political Economy and Quantitative Methods. University of Pavia.*

Perrakis, K., Ntzoufras, I. and Tsionas, E.G. (2014). On the use of marginal posteriors in marginal likelihood estimation via importance-sampling. *Computational Statistics & Data Analysis* **77**, 54-69.

Raftery, A. (1988). Approximate Bayes Factors for Generalized Linear Models. Technical Report **121**, Department of Statistics, University of Washington.

Raftery, A. (1996). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Model. *Biometrika* **2**(83), 251-266.

Rosenkranz, S.L. (1992). The Bayes Factor for Model Evaluation in a Hierarchical Poisson for Area Counts. *Unpublished Ph.D. dissertation, Department*

of Biostatistics, University of Washington.

Rousseeuw, P. J., and van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points (with discussion). *Journal of the American Statistical Association* **85**, 633-651.

Xie, W., Lewis, P., Fan, Y., Kuo, L., and Chen, M. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology* **60**, 150?160.

Greek Statistical Institute
Proceedings of the 30th Panhellenic Statistics Conference (2017), pp.109-123

TIME SERIES WITH INTERDEPENDENT LEVEL AND SECOND MOMENT: TESTING; CONSEQUENCES FOR MODELLING; PRELIMINARY RESULTS

Milionis E. Alexandros[1], Galanopoulos G. Nikolaos[2]
[1]Bank of Greece and Department of Mathematics, University of the Aegean
amilionis@bankofgreece.gr
[2]Department of Economics, National and Kapodistrian University of Athens
ngalano@econ.uoa.gr

SUMMARY

This work aims to fill an existing gap in the literature regarding the statistical testing for the existence and the identification of the character of time-varying second moment in its dependence on a non-constant mean level in time series. To this end we introduce a new statistical testing procedure with some considerable advantages over the existing ones. Amongst others we argue that the existing statistical tests are insufficient and sometimes lead to biased results. Further the effect of the application of this methodology on some crucial elements of time series modelling such as forecasting, seasonal adjustment and outlier detection is examined, through case studies conducted on a comparative basis using both the new methodology and an established one. Our data set comprises time series on monthly external trade statistics for Greece. The particular data were selected owing to their obvious importance given the continuing economic crisis in that country. The resulted empirical evidence is in favor of the new approach.

Key Words: applied time series analysis, seasonal adjustment, forecasting, time series transformations, detection of outliers, Greek external trade time series.

1. INTRODUCTION

Over the last five decades a vast volume of research work, in both theoretical and applied level, has been devoted on time series with time-varying second moment. This non-constancy in the second moment may be due to various reasons. For the purposes of this work it is methodologically useful to distinguish between type (i): series with conditionally non constant, but unconditionally constant variance, and type (ii): series with non-constant variance both conditionally and unconditionally. This work focuses mainly on the latter.

If variance is functionally related to the level and the latter is non-stationary (which most often is indeed the case), the variance is not constant both conditionally and

unconditionally [this is the typical case of type (ii)]. Hence, the process is non-homogeneously non-stationary in the sense of Box and Jenkins (1976) and cannot be made stationary by simply differencing. One way to tackle variance non-stationarity is to employ the well-known class of the so-called Box and Cox transformations (Box and Cox, 1964) given by:

$$f(y_t) = \begin{cases} y_t^{\lambda} \ if \ \lambda > 0 \\ \log y_t \ if \ \lambda = 0 \\ -y_t^{\lambda} \ if \ \lambda < 0 \end{cases} \quad (1)$$

In spite of its importance for time series modelling, at the theoretical level there is not much work on the detection and correction of non-constancy in the variance owing to its dependence on a non-stationary mean level. Further, at the practical level the treatment of non-stationary variance is not only insufficient (indeed, when a particular Box-Cox transformation is used its selection is often arbitrary) but also, occasionally, biased towards over-rejection of the null hypothesis of unconditionally constant variance, as we argue later on in this work.

The existing statistical approaches for detection and correction of variance non-stationarity appear to have several disadvantages as: (i) they detect variance non-stationarity, but the correction they suggest is not formally and rigorously documented (e.g. Hay and McLeay 1979; Milionis and Davies, 1994); (ii) usually suffer from subjectivity (see for instance Mills, 1990 for a short review); and (iii) although they may do detect variance non stationarity and are formally suggestive for a solution, they lack robustness (Milionis, 2004; Milionis, 2003).

The aim of this work is to introduce a formal econometric approach, which not only allows the detection of non-stationary variance and is suggestive of the transformation necessary to correct for it, but also is robust to the particular partition of a time series –a procedure necessary for the test- and the possible existence of outliers. Further, the possible advantages of the application of this methodology on some crucial elements of time series modelling such as forecasting, seasonal adjustment and outlier detection, as compared to existing methods, are examined.

2. DATA

Data set comprises the time series on monthly external trade statistics from the Balance of Payments for Greece. The particular data were selected due to their obvious importance given the continuing economic crisis in the country and the large current account deficit of Greece at the beginning of the economic crisis. This current account deficit is attributed primarily to the deficit of the balance of Goods (see press releases at the web site of the Bank of Greece). It is apparent that proper statistical modelling is vital for the short-term monitoring and forecasting of such series. The

Table 1. *Dates of important events during data period*

Event	Date
Lehman Brothers' bankruptcy	15/09/2008
Commencement of the first economic adjustment programme for Greece	06/05/2010
Commencement of the second economic adjustment programme for Greece	13/02/2012

Figure 1. *Imports of Goods (in million euro)*

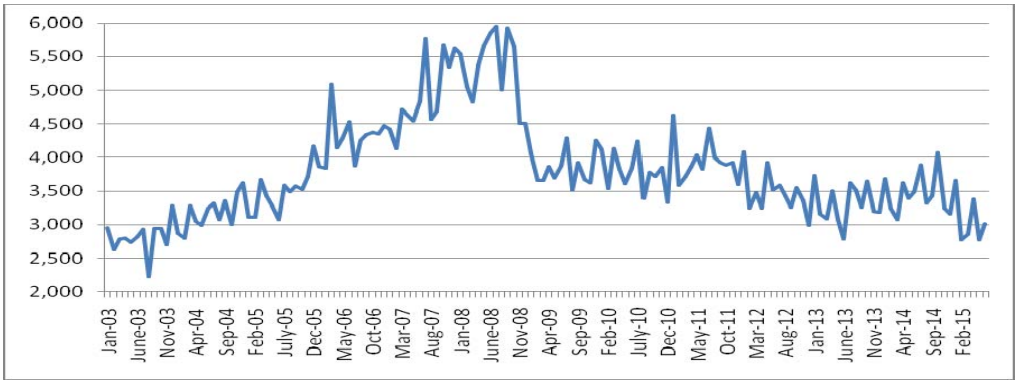
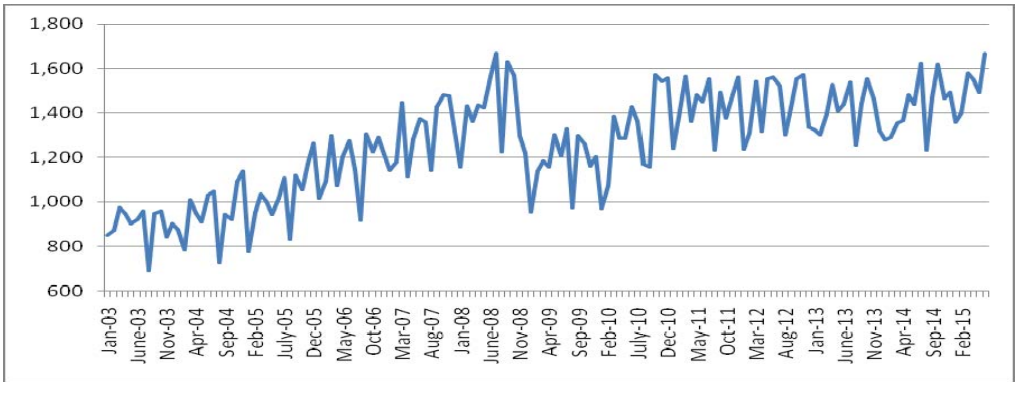


Figure 2. *Exports of Goods without fuels and ships (in million euro)*



data cover the period from January 2003 to June 2015 and consist of one hundred and fifty (150) monthly observations of Total Imports and Total Exports of Goods excluding fuels and ships (source: Bank of Greece). The dates of some of the important events occurred during the crisis time period are quoted in Table 1, while a graphical representation of the time series are shown in Figures 1 and 2.

The sharp decline in both series at end 2008, which is conspicuous from the visual inspection of those figures, may be attributed, amongst others, to the Lehman Brothers' bankruptcy and the subsequent sharp reduction in economic activity.

3. STATISTICAL TESTING APPROACH

3.1 Description

As in most other similar studies (Mills, 1990; Milionis and Davies 1994; Milionis 2004), for the statistical testing approach used in this work time series are partitioned into segments (subsamples) of equal length. For each subsample the (local) mean (LM) as well as the (local) standard deviation (LSD) are calculated. Local standard deviation is assumed to be functionally dependent on Local Mean in a non-linear fashion as follows:

$$LSD = \alpha LM^{\beta} e^{u}, \quad (2)$$

where α, β are model parameters, e is the base of natural logarithms and u the stochastic disturbance. Model parameters α, β are estimated via Ordinary Least Squares (henceforth OLS) using the corresponding log-log model. The estimated value of β ($\hat{\beta}$) provides the necessary information for the type of data transformation needed to ensure variance stationarity (e.g. for the most popular transformations, namely the log-transformation and the square root one, the log transformation corresponds to β =1, and the square root transformation corresponds to β =0.5). To ensure robustness with respect to the particular partition and the possible existence of outliers the procedure is repeated for different partitions. The number of different partitions is at least equal to the number of divisors of the series' length, giving quotient (series length over divisor) ≥ 5 and restricting the size of subsamples to be ≥ 5.

3.2 Notation, Equations, Statistical Hypotheses and comments

3.2.1 Notation

Before the description of the testing procedure some explanation on the notation and definition of the various symbols is necessary.

Index (k) indicates the ascending number of a subsample in a partition;
index (j) indicates the ascending number of the particular partition, $j=1,2,\ldots,j_{max}$;
index i_j represents the maximum value of k (number of subsamples) in partition j;
N is the total length (size) of the initial time series;
n_{ij} represents the size of subsamples in partition j;
$\hat{\beta}_j$ is the estimate of the exponent β using subsamples derived from partition with ascending number j;
An asterisk (*) over a symbol denotes the corresponding transformed data, or the corresponding estimate derived from the transformed data;
$\hat{u}_{jk}$, $\hat{\varepsilon}_j$, $\hat{u}^*_{jk}$ are independent of each other regression residuals;
$i_j = (N/n_{ij})$, if (N/n_{ij}) is an integer; $n_{ij}>5$;
$i_j = int(N/n_{ij}) + 1$, if (N/n_{ij}) is not an integer; $n_{ij}>5$

3.2.2 Statistical testing procedure

The statistical testing procedure comprises three stages as follows:

$\hat{\beta}_j$ is estimated for each partition j, j=1,2,…, j_{max} via OLS from the model (**First stage):**

$$\ln(LSD_{jk}) = \ln(\alpha_j) + \hat{\beta}_j \ln(LM_{jk}) + \hat{u}_{jk}, \quad (3)$$

$\hat{\beta}$ is estimated via OLS from the model (**Second Stage**):

$$\hat{\beta}_j = \hat{\beta} + \hat{d}j + \hat{\varepsilon}_j, \quad (4)$$

Model using the transformed data (**Third Stage**):

$$\ln(LSD^*_{jk}) = \ln(\alpha^*_j) + \hat{\beta}^*_j \ln(LM^*_{jk}) + \hat{u}^*_{jk}. \quad (5)$$

3.2.3 Statistical Hypotheses and comments

Applying the procedure described above, it can be made possible to state and test the following statistical hypotheses:

1) **H_a:** β_j**=0** $\forall$ **j** (or at least the majority of β_j s,).

   This hypothesis can be tested from the first stage and is utilized to ensure that indeed there exists a dependence of local standard deviation on local mean. Failure to reject **H_a** means that there is no such dependence and therefore, the algorithm stops.

2) **H_b:** d **=0 (Robustness test)**

   The dependent variable in Eq. 4 (second stage) is the estimate of β derived from the partition of ascending number j ($\hat{\beta}_j$), while the independent variable is the ascending number of the partition itself (j). Therefore, $\hat{d}$ is the estimate of the slope of the regression. This hypothesis states that the slope $\hat{d}$ should not be statistically significant, and non-rejection of it, means that $\hat{\beta}$ is robust to any particular partition of the series, or outliers. Additionally, non-rejection of **H_b** also ensures a better estimate of β by making more efficient use of information available in all partitions.

3) **H_c:** β^*_j**=0** $\forall$ **j (Under-transformation test)**

   β^*_j are the corresponding β_j for the transformed data estimated from the third stage. This test states that there is no remaining dependence of local mean on local standard deviation in the transformed data. Hence, non rejection of this hypothesis ensures that the chosen transformation has adequately rendered an unconditionally stable variance.

4. RESULTS-DISCUSSION

The effect of the application of this methodology on some crucial elements of time series modelling such as forecasting, seasonal adjustment and outlier detection is examined, through a comparative empirical analysis using both the new methodology and an established algorithm used in the widely used software TSW. TSW stands for TRAMO-SEATS for Windows, a Windows version of the programmes TRAMO and SEATS (see Gómez and Maravall, 1996). Once we apply the new testing procedure and a proper data transformation is selected, we will also use TSW for further analysis.

Outliers are automatically detected, classified and corrected using the Chen and Liu (1993) approach. Three types of outliers are detected according to their effect in a time series:

- Additive outliers (AO), which affect only a single observation of the series,
- Transitory Change outliers (TC), the effect of which is not extinguished in the next observation, as is the case with the additive outliers, but damps out gradually over the subsequent few periods, and
- Level shifts (LS), which imply a step change in the level of the series.

Statistical forecasts are made after the series are "linearized" according to the following model:

$$y_t = w_t^{'}\beta + C_t^{'}\eta + \sum_{j=1}^{k} \alpha_j \lambda_j(B) I_t(t_j) + x_t \qquad (6)$$

where $\beta = (\beta_1,...,\beta_n)$, is a vector of regression coefficients, $w_t^{'} = (w_{1t},...,w_{nt})$ denotes n regression or intervention variables, $C_t^{'}$ denotes the matrix with columns possible calendar effect variables (e.g. trading day) and η the vector of associated coefficients, $I_t(t_j)$ is an indicator variable for the possible presence of an outlier at period t_j, $\lambda_j(B)$ captures the transmission of the j-th effect (and α_j denotes the coefficient of the outlier in the multiple regression model with k outliers. From the covariates implied in eq. 6 in the present case the most influential are the outliers of various types (further details and estimation results are provided in section 4). Finally, x_t follows in general the multiplicative ARIMA(p,d,q)(P,D,Q)$_s$ model:

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D x_t = \theta(B)\Theta(B^s)\varepsilon_t \qquad (7)$$

where:

- $\phi(B) = 1 - \varphi_1 B - ... - \varphi_p B^p$ is the so–called autoregressive polynomial of order p;

Table 2. *Estimates of* $\hat{\beta}_j$ *for the various partitions for the first stage regression*

Subsample size (n_{ij})	5	6	8	10	12	14	16	18	20
Number of subsample pairs (i_j)	30	25	19	15	13	11	10	9	8
$\hat{\beta}_j$	1.058	1.203	1.052	0.863	1.090	0.955	1.032	0.791	1.246
t-statistic	2.642	3.230	3.211	2.096	4.204	2.297	3.321	2.278	2.758
p-value	0.013	0.004	0.005	0.056	0.001	0.047	0.010	0.057	0.033

- $\theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$ is the so–called moving average polynomial of order q;
- $\nabla^d \equiv (1-B)^d$ is the arithmetic difference operator of order d;
- $\nabla_s^D \equiv (1-B)_s^D$ is seasonal arithmetic difference operator of order D and seasonality s;
- $\Phi(B^s) = 1 - \Phi_1 B^s - ... - \Phi_P B^{P \cdot s}$ is the so–called seasonal autoregressive polynomial of order P and seasonality s;
- $\Theta(B^s) = 1 - \Theta_1 B^s - ... - \Theta_Q B^{Q \cdot s}$ is the so–called moving average polynomial of order Q and seasonality s;
- ε_t is the stochastic disturbance.

Seasonal adjustment is based on ARIMA model-based signal extraction. This method uses the Burman-Wilson algorithm (Burman, 1980) to decompose a time series in unobserved components [for further details see Maravall (1995), Gómez and Maravall (1996)].

4.1 Comparative analysis of the time series of "Imports of goods"

4.1.1 Analysis using of the new statistical testing approach

The results for the first stage regressions for the time series "Imports of goods" are presented in Table 2. It can be seen that in all subsample pairs the $\hat{\beta}_j$ estimates are statistically significant at 10% significance level and in almost all subsample pairs at 5% significance level (exceptions only for the partitions with subsample size 10 and 18 where the estimates are "marginally" significant for the 5% significance level). Hence, $\mathbf{H_a}$ is clearly rejected.

Figure 3. *Graphical representation of the first stage regressions results*

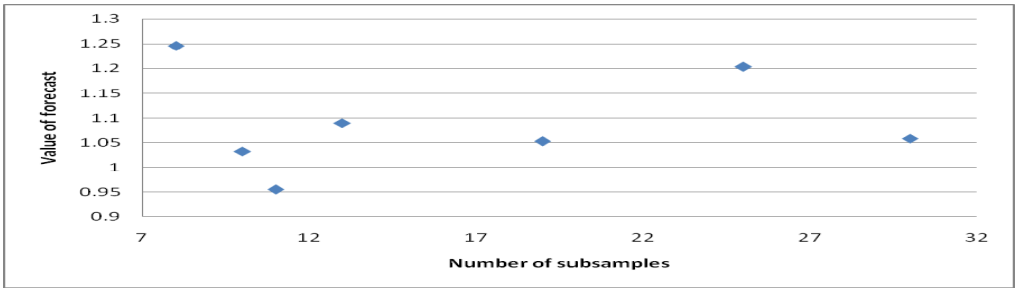


Table 3. *Results for the second stage regression*

	estimate	t-statistic	standard error
$\hat{\beta}$	1.082	10.956	0.099
$\hat{d}$	0.0005	0.101	0.005

Table 4. *Results for "under-transformation" test*

Subsample size (n_{ij})	5	6	8	10	12	14	16	18	20
Number of subsample pairs (i_j)	30	25	19	15	13	11	10	9	8
$\hat{\beta}_j^*$	0.305	1.655	0.222	-1.59	0.558	-0.75	5.882	-1.61	6.470
t-statistic	0.092	0.540	0.085	-0.50	0.258	-0.23	1.325	-0.56	1.350
p-value	0.927	0.594	0.933	0.623	0.801	0.821	0.222	0.592	0.220

The results for the first stage regressions are also depicted in Figure 3, where the x-axis represents the number of subsamples and the y-axis the value of exponent $\hat{\beta}_j$.

From the visual inspection of Figure 3 it is apparent that no systematic association between $\hat{\beta}_j$ estimates and the sample size seems to exist.

This is further supported by the results of the second stage regression by which $\mathbf{H_b}$**:** d **=0** can be formally tested**.** Those results are presented in Table 3. As is evident, the constant β is statistically significant at the 5% level and equal to 1.082, whereas the slope d is not statistically significant. Hence, $\mathbf{H_b}$ cannot be rejected.

The above results clearly suggest that the original data are variance non stationary and should be log-transformed.

To examine, whether or not, the chosen logarithmic transformation is indeed sufficient to stabilize the variance the so-called "under-transformation" test is further performed. To this end the logarithms of the original data are subjected to the logarithmic transformation once more and the parameters of the third stage regression are estimated.

The results of the "under-transformation" test are presented in Table 4.

As is evident from the results of Table 4 all $\hat{\beta}_j^*$ are not statistically significant, therefore the hypothesis $\mathbf{H_c}$**:** $\beta_j^* = \mathbf{0} \; \forall \; \mathbf{j}$ is not rejected.

4.1.2 Analysis using exclusively the TSW testing approach

The same series was reanalyzed following the standard TSW procedure. The way TSW tests, whether or not, the data need to be transformed in order to stabilize the variance is based on a variant of the so-called range-mean regression (see Gomez and Maravall, 1996). More specifically, the series is divided in subsamples and the range and mean for each subsample are calculated. Then a regression model using the subsamples' ranges and means is estimated. If the regression slope is found to be significant the data are log-transformed.

Using the TSW procedure, TSW also suggested the logarithmic transformation of the original data, as was the conclusion using the new approach. However, when the TSW procedure was repeated once again with the log-transformed data ("under-transformation" test), TSW suggested a logarithmic transformation again(!). Indeed, TSW output states that:

«*LOG-LEVEL PRETEST: SSlevels/(SSlog*Gmean(levels)^2)=1.0078847 LOGS ARE SELECTED*».

Therefore, TSW seems to be biased towards rejection of the null hypothesis of no transformation.

4.1.3. The effect of data transformation on univariate modelling and outlier detection

It is of much interest to further investigate how variance non-stationarity could potentially affect the specification of the univariate ARIMA model and the detection of outliers. Table 5 below presents the results on univariate ARIMA modelling with and without the log-transformation, while the estimation details are quoted in the Appendix. From the results of Table 5 it is apparent that when variance non-stationarity is taken into account the univariate model is the so-called "airline" model, often encountered in data with seasonality (see Box and Jenkins, 1976). In contrast,

Table 5. *Univariate ARIMA modelling*

ARIMA model with linearized original data [Model (1)]
ARIMA(2,1,0) $(1,0,1)_{12}$ MODEL SPECIFICATION
$(1-0.583B-0.401B^2)(1+0.857B^{12})(1-B)Y_t = (1+0.594B^{12})\varepsilon_t$
ARIMA model with linearized transformed data (TSW and new method) [Model (2)]
ARIMA(0,1,1) $(0,1,1)_{12}$ MODEL SPECIFICATION
$(1-B)(1-B^{12})\log Y_t = (1+0.609B)(1+0.592B^{12})\varepsilon_t$

Table 6. *Outlier Detection (series of Imports of goods)*

Outliers with original data	Outliers with log-transformed data (TSW, New Method)
39 TC (3 2006), 71 LS (11 2008), 97 A0 (1 2011)	72 LS (12 2008), 96 A0 (12 2010)

without taking into account variance non stationarity a much more complicated ARIMA model is selected. Hence, the presence of variance non stationarity leads to seriously misspecified univariate ARIMA models, a result that is in accordance to that of Milionis (2004).

The results on the detection of outliers are presented in Table 6. As is evident a TC outlier at period 39 in the original data does not exist in the transformed data. Further, the LS outlier at period 71 (November, 2008) in the original data, which is most likely related to the Lehman's bankruptcy, has shifted forward one period in the transformed data, while the AO outlier in period 97 in the original data has shifted backward one period in the transformed data. Hence, in properly transformed data the pattern of detected outliers is clearly different, a conclusion that is also in accordance with that of Milionis (2004).

4.2 Comparative analysis of the time series of "Exports of goods excluding fuels and ships"

4.2.1 Testing for variance non-stationarity

Table 7 presents the results of the first stage regressions for the series of Exports of goods excluding fuels and ships. From these results it is evident that $\hat{\beta}_j$ is not statistically significant in the 5% significance level, except for the partition with

Table 7. *Estimates of* $\hat{\beta}_j$ *for the various partitions for the first stage regression*

Subsample size (n_{ij})	5	6	8	10	12	14	16	18	20
Number of subsample pairs (i_j)	30	25	19	15	13	11	10	9	8
$\hat{\beta}_j$	0.654	0.723	0.530	0.467	0.455	0.440	0.595	0.466	0.246
t-statistic	1.812	2.110	2.015	1.713	1.233	1.111	1.167	1.279	0.392
p-value	0.081	0.046	0.060	0.110	0.243	0.295	0.277	0.241	0.709

Table 8. *Univariate ARIMA modelling for the series of Exports of goods excluding fuels and ships*

ARIMA model with linearized original data [Model (3)]
ARIMA(0,1,1) $(0,1,1)_{12}$ MODEL SPECIFICATION
$(1-B)(1-B^{12})Y_t = (1+0.584B)(1+0.7.64B^{12})\varepsilon_t$
ARIMA model with linearized transformed data (TSW and new method) [Model (4)]
ARIMA(0,1,1) $(0,1,1)_{12}$ MODEL SPECIFICATION
$(1-B)(1-B^{12})\log Y_t = (1+0.566B)(1+0.821B^{12})\varepsilon_t$

subsample size 6. Thus, according to the new testing approach in this case the series variance is (unconditionally) stationary and no transformation of the original data is required. However, the conclusion is different when the approach of TSW is followed, as TSW log-transforms the data as a consequence of the range-mean regression. Indeed, TSW output states that : «*LOG-LEVEL PRETEST: SSlevels/(SSlog*Gmean(levels)^2)=1.0106056 LOGS ARE SELECTED*».
Therefore, once again, TSW is biased towards the logarithmic transformation, whereas no transformation of the original data needs to be performed.

4.2.2 Further Analysis

Table 8 below presents the results on univariate ARIMA modelling with and without the log-transformation for the series of Exports of goods excluding fuels and ships, while the estimation details are quoted in the Appendix. From the results of Table 8 it is apparent that in contrast to the univariate models referring to the series of Imports of Goods here the differences in the two univariate models are of minor character, as

Table 9. *Outlier Detection (series of Exports of goods excluding fuels and ships*)

Outliers with original data (new method)	Outliers with log-transformed (TSW)
71 LS (11 2008), 93 AO (9 2010)	72 LS (12 2008), 93 AO (9 2010)

Table 10. *Values of Mean absolute percentage error of forecasts*

	MAPE (%)	
Forecasts	Original Data (New Method)	Log-transformed data (TSW)
Twelve step ahead	3.618	4.501
One step ahead	3.265	3.778

Table 11. *Differences in Seasonally Adjusted Series produced from original data versus transformed data*

MAPE (%)	Minimum Percentage Error (%)	Maximum Percentage Error (%)
1.297	-4.314	5.515

in both cases the univariate model is of the same type i.e. the so-called "airline" model. The differences are confined only to the estimated values of the parameters of the two models. This is not surprising as with the series of Exports of goods excluding fuels and ships both the original and the log-transformed series are variance stationary. Indeed this result advocates our previous conclusion for the series of Exports of goods where the pronounced difference in the character of the univariate ARIMA model for the original and the log-transformed data was attributed to the existence of non-stationary variance in the original data series.

The results on the detection of outliers with both the original and the log-transformed data for the series of Exports of goods excluding fuels and ships are quoted in Table 9. As is evident the AO outlier is the same in both cases, while the level shift, which as mentioned earlier is related to Lehman's bankruptcy has only been moved forward by one time period in the log-transformed data.

Our next task is to examine, whether or not, the data transformation affects the forecasting performance of the univariate models, as well as the seasonally adjusted series. The former was evaluated using the mean absolute percentage forecast error.

Table 10 presents the results. As is evident both one-step-ahead and twelve-step-ahead forecasts with no transformation (as suggested by the new method) are superior in terms of the MAPE value, as compared to the corresponding forecasts with the data log-transformed, as suggested by TSW.

At this point the provisional character of the result regarding the forecasting performance should be stressed, especially in view of the fact that thus far in the literature the empirical evidence on the relation of the forecasting performance and the chosen data-transformation is mixed (Mills, 1991, Makridakis et. al, 1998, Nelson and Granger, 1979). Further analysis using several series is needed to provide additinal evidence in support of this argument.

Finally the MAPE statistic was also employed to assess the differences in the seasonally adjusted series produced from original data versus transformed data. The results are presented in Table 11. It is remarked that from the results of Table 11 substantial differences are observed, as the MAPE between the two seasonally adjusted series is approximately 1.3%, with the minimum percentage error to be equal to -4.3% and the maximum percentage error to be equal to 5.5%.

5. CONCLUSIONS

In this work a new statistical testing procedure for variance non-stationary time series is proposed. This procedure improves the existing ones at it combines detection, correction and robustness. In addition it was shown empirically that the existing tests, such as the one in the widely used algorithm of TSW software, provide biased results. Further, it is argued that the type of data transformation and the entailed correction for variance–non stationarity is crucial for the detection of outliers and the seasonal adjustment of the original time series. In addition, the empirical results provide evidence of an improved forecasting performance by the proper use of a data transformation, a result that needs further backing by additional empirical evidence. Overall, the proposed statistical testing procedure, placed in a more general framework, seems to be a promising tool in applied time series analysis.

APPENDIX

Parameter estimation of Univariate ARIMA models

Table A.1*Parameter estimation of model (1)*

Parameter	Value	t-statistic
ϕ_1	0.583	7.480
φ_2	0.401	5.290
Φ_1	-0.857	-10.890
Θ_1	-0.594	-4.810

Table A.2. *Parameter estimation of model (2)*

Parameter	Value	t-statistic
θ_1	-0.609	-8.750
Θ_1	-0.592	-8.360

Table A.3. *Parameter estimation of model (3)*

Parameter	Value	t-statistic
θ_1	-0.584	-8.190
Θ_1	-0.764	-13.490

Table A.4. *Parameter estimation of model (4)*

Parameter	Value	t-statistic
θ_1	-0.566	-7.850
Θ_1	-0.821	-16.430

ΠΕΡΙΛΗΨΗ

Πρωταρχικός σκοπός της παρούσας εργασίας είναι να καλύψει ένα υπάρχον κενό στη βιβλιογραφία που αφορά τον στατιστικό έλεγχο για την ύπαρξη και τον προσδιορισμό του χαρακτήρα της χρονικής μεταβολής της δεύτερης ροπής όταν εξαρτάται από ένα μη σταθερό μέσο επίπεδο σε χρονοσειρές. Για τον σκοπό αυτό προτείνεται μια νέα διαδικασία ελέγχου με σημαντικά πλεονεκτήματα σε σχέση με τις υπάρχουσες διαδικασίες. Όπως επιχειρηματολογείται, οι υπάρχοντες στατιστικοί έλεγχοι είναι, μεταξύ άλλων, ανεπαρκείς και μερικές φορές οδηγούν σε μεροληπτικά συμπεράσματα. Περαιτέρω, η επίδραση της εφαρμογής αυτής της μεθοδολογίας σε μερικά βασικά στοιχεία της μοντελοποίησης χρονοσειρών, όπως είναι η πρόβλεψη, η εποχική διόρθωση και η ανίχνευση ακραίων τιμών, εξετάζεται μέσω συγκριτικής μελέτης περιπτώσεων χρησιμοποιώντας τόσο τη νέα μεθοδολογία όσο και την καθιερωμένη. Τα δεδομένα που χρησιμοποιήθηκαν είναι μηνιαίες χρονοσειρές από

το εμπορικό ισοζύγιο (συνολικές εισαγωγές –εξαγωγές αγαθών) για την Ελλάδα. Τα συγκεκριμένα δεδομένα επελέγησαν λόγω της προφανούς σημασίας τους, δεδομένης της συνεχιζόμενης οικονομικής κρίσης στη χώρα. Τα εμπειρικά ευρήματα που προέκυψαν είναι υπέρ της νέας μεθοδολογίας η οποία ιδωμένη κάτω από ένα γενικότερο πλαίσιο δύναται να αποτελέσει ένα χρήσιμο μεθοδολογικό εργαλείο στην εφαρμοσμένη ανάλυση χρονοσειρών.

REFERENCES

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, Revised Edition, San Francisco: Holden-Day.

Box, G. E. P. and Cox, D. R. (1964). 'An Analysis of Transformations', *Journal of the Royal Statistical Society*, **26**, 211-243.

Burman, J. P. (1980). 'Seasonal Adjustment by Signal Extraction', *Journal of the Royal Statistical Society*, **143**, 321-337.

Chen, C. and Liu, L.-M. (1993). Forecasting Time Series With Outliers, *Journal of Forecasting*, **12**, 13-35.

Gomez, V. and Maravall, A. (1996). Programmes SEATS and TRAMO: instructions for the User. *Working Paper No 9628, Bank of Spain*.

Hay, A. R. and McCleary, R. (1979). Box-Tiao Time Series Models for Impact Assessment, *Evaluation Review*, **3**, 277-314.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (1998). *Forecasting Methods and Applications*, 3rd ed., USA, NJ, Wiley.

Maravall, A. (1995). *Unobserved components in economic time series*, In Handbook of Applied Econometrics (Eds) M. H. Pesaran and M. Wickens, Blackwell, Oxford.

Milionis, A. E. and Davies, T. D. (1994). Regression and Stochastic models for Air Pollution, Part II, *Atmospheric Environment*, **28**, 2801-2810.

Milionis, A. E. (2003). Modelling Economic Time Series in the Presence of Variance non-Stationarity: a Practical Approach, *Working Paper No 8, Bank of Greece*.

Milionis, A. E. (2004). The importance of Variance Stationarity in Economic Time Series Modelling; a Practical Approach, *Applied Financial Economics*, **14**, 265-278.

Mills, T. C. (1990). *Time Series Techniques for Economists*, Cambridge University Press, Cambridge.

Nelson, H. L. and Granger, C. W. L. (1979). Experience with using the Box-Cox transformation when forecasting economic time series. *Journal of Econometrics*, **10**, 57-69.

Greek Statistical Institute
Proceedings of the 30th Panhellenic Statistics Conference (2017), pp.124-131

MODELING AND PROJECTING HEAT-RELATED MORTALITY

Tsangari Haritini
University of Nicosia
tsangari.h@unic.ac.cy

ABSTRACT

Empirical evidence has shown that the temperature-mortality relation exhibits both non-linearities and lag effects. The present study aims to model and project the effect of extreme weather on mortality, using real data from Cyprus. The modeling procedure manages to capture both non-linearities and delayed effects simultaneously. A temperature function is first created within the framework of Distributed Lag Non-linear Models. The temperature function is then incorporated into a Generalized Linear Model, with quasi-Poisson regression to allow for overdispersion, adjusting for possible confounders. All the results regarding the effect of heat on mortality together with projections are presented and discussed.

Keywords: Non-linearities; lag effects; mortality; climate change; Cyprus.

1. INTRODUCTION

The world is currently experiencing climate change in the form of increased occurrence of heat waves and large temperature fluctuations, potentially leading to increased mortality worldwide [e.g,, Kovats and Hajat (2008); Gosling et al. (2009)].

The projected global temperature rise is likely to be from 1.8-4.0°C over the 21st century (IPCC, 2013). The largest warming is likely to be over southern Europe and the Mediterranean, with annual mean temperature increases as large as +5.5°C [Zahariadis (2012)].

Related literature has shown consistent evidence of association between high temperatures and mortality [e.g., Baccini et al. (2008); Almeida et al. (2010); Zanobetti and Schwartz (2008)]. Therefore, heat-related mortality is causing great public health concern in many countries.

The aim of the current paper is to model and quantify the effect of extreme weather on mortality, using a methodology that manages to capture simultaneously all the

characteristics of the temperature-mortality relation. The modeling procedure will be applied in data from Cyprus, which has a typical Mediterranean climate.

2. THE MODELING PROCEDURE

2.1 Theoretical framework

Empirical evidence from various studies has shown that the temperature-mortality relation exhibits two main characteristics: first, non-linearities, where we have higher mortality at temperature extremes, and, second, "delayed" or "lag" effects, where high temperatures have an effect on mortality not only on the same day, but on the next days as well [Armstrong (2006); Braga et al. (2001); Anderson and Bell (2009); Gasparrini et al. (2010)]. Therefore, an efficient methodology is required, which can capture both these characteristics under one modeling framework.

In addition, the relationship between temperature and mortality may be confounded by measured or unmeasured confounders, including meteorological indicators, long- and short-term seasonality, as well as air pollution [Armstrong et al. (2011); Dominici et al. (2000); Iñiguez et al. (2010)].

2.2 The data

The study area was Cyprus, an island with a Mediterranean climate. The analysis concentrated in the warm periods of each year, namely April to September, with data available for the years 2004-2009. The climate during these months is characterized by high temperatures, negligible rainfall and some isolated thunderstorms.

Daily meteorological data were collected by the Cyprus Meteorological Service, in the five main urban centers of the island, namely Athalassa (Nicosia), Larnaca, Limassol, Paralimni and Paphos. Daily mortality data were provided by the Ministry of Health of the Republic of Cyprus. Particulate matter (PM_{10}) was selected as the variable for air pollution, where daily data for particulate matter were obtained from the Department of Labour Inspection at the Ministry of Labour and Social Insurance: daily levels of PM_{10} were taken from the records of 5 stations (Agia Marina Xyliatou, Nicosia, Larnaca, Limassol and Paphos). Cyprus was considered as a total area, using the combined data from all the stations.

2.3 The model

A Generalized Linear Model (GLM) was used for the mortality count data, with quasi-Poisson regression to allow for overdispersion. The general form of the model for the mortality counts, Y_t, t=1,….n, is given in equation (1):

$$g(\mu_t) = \alpha + \sum_{j=1}^{J} s_j(x_{tj}; \boldsymbol{\beta}_j) + \sum_{k=1}^{K} \gamma_k u_{tk}\,, \qquad (1)$$

where $\mu = \mathrm{E}(Y)$, g is a monotonic link function, with a distribution from an exponential family, and the functions s_j denote smoothed relationships between the variables x_j and the linear predictor, defined by the (unknown) parameter vectors $\boldsymbol{\beta}_j$. More specifically, s_1 is the temperature function, s_2 is the function for the meteorological indicator "relative humidity" and s_3 is the function for long-term trend. The variables u_k in equation (1) include other predictors with linear effects specified by the related coefficients, γ_k, such as short-term seasonality, as well as air pollution, which is the average of lags 0 and 1.

Observing the non-linear shape of the temperature-mortality relationship for the Cyprus dataset, the smooth function s_1 was created based on the recently developed methodology of Distributed Lag Non-Linear Models (DLNM). Since the temperature-mortality relation exhibited two main characteristics, non-linearities and "lag" effects, both these dimensions had to be captured. The methodology involved the formulation of the appropriate "cross-basis", a bi-dimensional space of functions describing simultaneously the non-linear shape of the relationship and the distributed lag effects [Gasparrini et al. (2010)]. More specifically, choosing a cross-basis amounts to specifying two independent sets of basis functions, one for each dimension. A "basis" is a space of functions used to define the relationship. The choice of the basis involves the related basis functions, completely known transformations of the original predictor generating a new set of transformed variables. These two independently chosen functions, one for the non-linear shape and one for the delayed effect, are then combined to generate cross-basis functions. Cross-basis functions are thus sums of products of the basis functions for temperature and lag.

First, we choose a basis for **x** to define the dependency in the space of the predictor. Then, we create the additional lag dimension for each one of the derived basis variables of **x**, with L being the maximum lag. This produces a $n \times v_x \times (L+1)$ array, which represents the lagged occurrences of each of the basis variables of **x**. Let **C** be an $(L+1) \times v_l$ matrix of basis variables for the lag vector $\boldsymbol{l}$ and let $\boldsymbol{\eta}$ be a vector of unknown parameters. Equation (2) shows the specification of the DLNM model function for temperature.

$$s(x_t; \boldsymbol{\eta}) = \sum_{j=1}^{v_x} \sum_{k=1}^{v_l} \mathbf{r}_{tj.}^{T} \boldsymbol{c}_{.k} \eta_{jk} = \boldsymbol{w}_{t.}^{T} \boldsymbol{\eta} \qquad (2)$$

where $\mathbf{r}_{tj.}$ is the vector of lagged exposures for the time t transformed through the basis function, $\boldsymbol{c}_{.k}$ is the k[th] column of matrix **C** and the vector $\boldsymbol{w}_{t.}$ is the t[th] row of the cross-basis matrix **W,** obtained by applying the cross-basis functions to x_t (we have $v_x \cdot v_l$ cross-basis functions).

For our data, the cross-basis for temperature comprises the "linear thresholds" function for the predictor and "strata" for the lag function, described in 1) and 2) as follows: 1) the "linear thresholds" model with a high threshold parameterization, k, was selected for the non-linearity dimension of the cross-basis, since after observing the data it was obvious that mortality is not related to temperatures below the threshold, while the effect is linear above it. This model can be represented by a truncated linear function $(x-k)+$ which equals $(x-k)$ when $x>k$ and 0 otherwise. 2) For the lag dimension of the cross-basis, a distributed lag model with strata constraints on the coefficients was chosen. In other words, specific cut-off points along the range of the predictor were applied, in order to define specific intervals, and then specify new variables through a dummy parameterization. More specifically, we assessed the effect of temperature on mortality with lags up to 10 days before the day of death. Three strata intervals were then defined, namely at lags 0-1, 2-5 and 6-10, with dummy parameterization assuming constant distributed lag effects along the strata levels. This constraint improved the precision of the estimates and avoided collinearity issues.

The abovementioned modeling procedure for the temperature-mortality relation has the advantage of capturing simultaneously both non-linearities and lag effects [Tsangari et al. (2016)]. The results of a DLNM can be interpreted by building a grid of predictions for each lag and for suitable values of the predictor, temperature. The relationship is summarized at single predictor or lag values, by cutting a "slice" of the grid along specific values. An estimate of the overall cumulative association can also be computed by summing all the contributions at different lags for each predictor value.

Regarding the other functions in equation (1), the function s_2, which controls for the non-linear effect of relative humidity, was found to be a natural cubic spline with 3 df, the function s_3 for long-term trends was a natural cubic spline with 4 df, while short-term seasonality was defined by dummy variables.

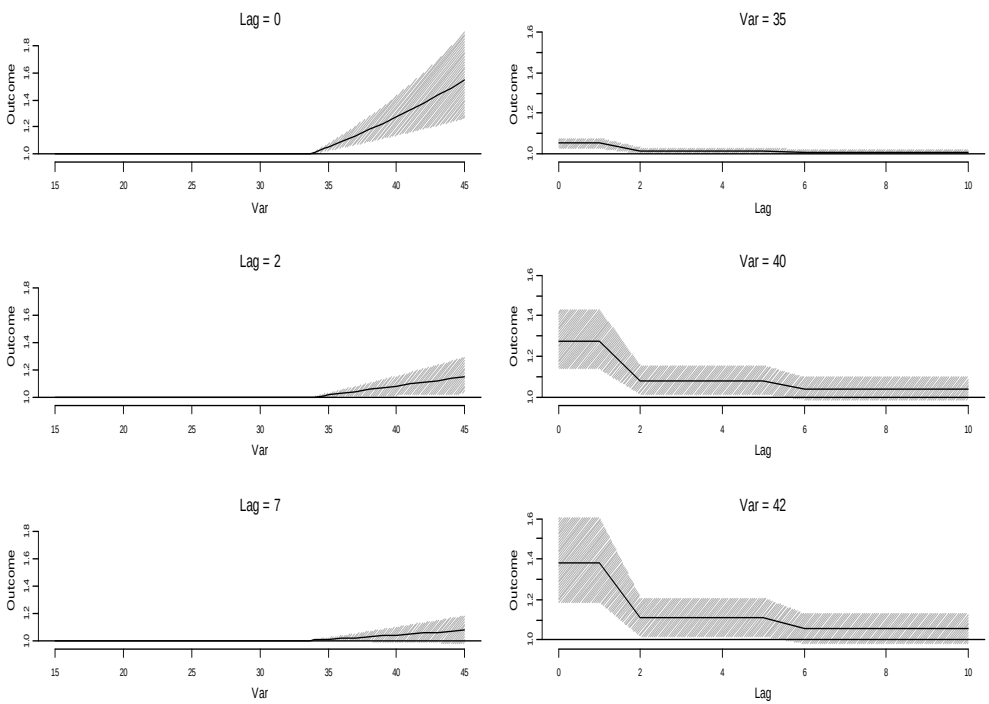
Finally, the potential effects of climate change on heat-related mortality in the future in Cyprus were estimated by calculating future mortality based on climate projections for the Mediterranean region, following a methodology applied in other regions of the world [see Hajat et al. (2014); Heaviside et al. (2016)]. The potential health impact was based on incremental increases in temperature above the data baseline period, from +1 °C to +5 °C, in line with projections of climate up to 2100.

3. RESULTS

In order to estimate the temperature function $s_{1,}$ the heat threshold, k, was identified first, based on the observed shape of the temperature-mortality relation. By testing a grid of temperatures from 31ºC to 35ºC in 0.1ºC increments and using model choice criteria, such as minimizing residual deviance and Akaike Information Criterion, the threshold was found to be 33.7 ºC. Therefore, the effect of temperature on mortality has zero slope up to 33.7 ºC and increases after this threshold.

The estimated increase in mortality for an increase of 1ºC above the threshold (i.e. the relative risk) was then estimated using the corresponding model of equation (1). High temperatures had a significant effect on mortality, even after adjusting for the effect of confounders. The relative risk for lags 0-1 was estimated to be 8.5%, while for lags 2-5 and 6-10 it was around 2% respectively, much lower compared to the effect during the same and next day of the event, which shows that the effect of heat on mortality was immediate. Figure 1 shows the risk of mortality for different lags (0, 2 and 7) and different temperatures (35^0C, 40^0C and 42^0C).

Figure 1. *Relative risk for various lags (0, 2 and 7) and temperatures (35^0C, 40^0C and 42^0C)*



Lags 0, 2 and 7 were chosen (graphs on the left of figure1) to demonstrate that the increase in effect for higher temperatures is more pronounced for lag 0, compared to lags 2 or 7. Temperatures 35^0C, 40^0C and 42^0C were selected (graphs on the right of figure 1) to demonstrate that the effect of heat is much more pronounced for higher temperatures (e.g. 42^0C compared to 35^0C or 40^0C).

Finally, estimation of potential effects indicated a significant number of heat-related deaths for current climatic conditions. Mortality rates appear to increase rapidly as temperatures rise. For example, for an increase of 1°C over the threshold temperature, heat-related mortality doubles, while for a 5 °C increase mortality is almost 800% the baseline.

4. CONCLUSION

The present study has used a methodology that captured the two characteristics of the mortality-temperature relation, non-linearities and distributed lag effects, simultaneously. The results showed that high temperatures result in increased mortality in Cyprus, independent of relative humidity, short- and long-term seasonality and air pollution.

It was shown that the effect is direct or immediate, since the risk is higher on the current and next day of a severe heat event and becomes lower during the following days, as we move further from the event. For example, the effect during the same or next day was around 4%, indicating that a 1^{o}C increase in maximum temperature above the hot threshold of 33.7^{o}C was associated with an estimated 4% increase in mortality during the same and next day, more than 4 times higher than the effect of each of the next few days. The results of a pronounced direct effect of heat (lags 0-1, compared with longer lags) on mortality agree with previous studies (e.g., Armstrong 2006; Braga et al.2001).

In addition to the immediate effect of heat, the results showed that the effect is more intense at extreme weather conditions: the increase in risk is higher at higher temperatures, but drops sharply after the first two days, as opposed to a more smooth effect for lower temperatures. For example, it is 40% higher at 42^0C, compared to 35^0C. The effect for temperatures of 35^0C and 37^0C is smoother, compared to the sudden drop for temperatures above 40^0C, as we move further away from the event (e.g. at lag 7 or one week after the event).

The adverse health effects of heat are largely preventable, if appropriate measures are implemented, including, among others, the setting up of early warning systems. Therefore, the results of the current study can be used as the research basis for linking

accurate forecasts of heat waves with effective preventive public health measures and interventions, targeting climatic variables.

ΠΕΡΙΛΗΨΗ

Εμπειρικά αποτελέσματα στη βιβλιογραφία έχουν δείξει ότι η σχέση μεταξύ θερμοκρασίας και θνησιμότητας παρουσιάζει μη-γραμμικότητα και χρονικές υστερήσεις. Η παρούσα έρευνα έχει στόχο να μοντελοποιήσει και να προβλέψει την επίδραση ακραίων καιρικών φαινομένων στη θνησιμότητα, χρησιμοποιώντας πραγματικά δεδομένα από την Κύπρο. Η διαδικασία μοντελοποίησης που χρησιμοποιείται επιτυγχάνει να χειριστεί ταυτόχρονα τη μη-γραμμικότητα και τα φαινόμενα χρονικών υστερήσεων. Πρώτα, η συνάρτηση για τη μοντελοποίηση της θερμοκρασίας δημιουργείται μέσα στα θεωρητικά πλαίσια Μη-Γραμμικών Μοντέλων Κατανεμημένης Υστέρησης (Distributed Lag Non-linear Models). Στη συνέχεια, η συνάρτηση θερμοκρασίας ενσωματώνεται μέσα σε ένα Γενικευμένο Γραμμικό Μοντέλο, με ημι-Poisson παλινδρόμηση ώστε να επιτρέπεται υπερσκέδαση, προσαρμόζοντας για πιθανούς συγχυστές. Όλα τα αποτελέσματα σχετικά με την επίδραση των υψηλών θερμοκρασιών στη θνησιμότητα και σχετικές προβλέψεις παρουσιάζονται και συζητούνται.

Acknowledgments: The project was co-financed by the European Regional Development Fund and the Republic of Cyprus, through the Research Promotion Foundation. The author would like to thank all the members of the CYPHEW group.

REFERENCES

Almeida, S.P., Casimiro, E., Calheiros, J. (2010). Effects of apparent temperature on daily mortality in Lisbon and Oporto, Portugal. *Environmental Health*, **9**, 1-7.

Anderson, B.G. and Bell, M.L. (2009). Weather-Related Mortality: How Heat, Cold, and Heat Waves Affect Mortality in the United States. *Epidemiology*, **20**, 205-213.

Armstrong, B. (2006). Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, **17**, 624-631.

Baccini, M., Biggeri, A., Accetta, G., Kosatsky, T., Katsouyanni, K., Analitis, A., Anderson, H.R., Bisanti, L., D'Ippoliti, D., Danova, J., Forsberg, B., Medina, S., Paldy, A., Rabczenko, D., Schindler, C., Michelozzi, P. (2008). Heat effects on mortality in 15 European cities. *Epidemiology*, **19**, 711-719.

Braga, L.F., Zanobetti, A., Schwartz, J. (2001). The Lag Structure between Particulate Air Pollution and respiratory and cardiovascular Deaths in 10 US Cities. *Journal of Occupational and Environmental Medicine*, **43**, 927-933.

Dominici, F., Samet, J.M., Zeger, S.L. (2000). Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society*, **163**, 263-302.

Gasparrini, A., Armstrong, B., Kenward, M.G. (2010). Distributed lag non-linear models. *Statistics in Medicine*, **29**, 2224–2234.

Gosling S.N., Lowe J. A., McGregor G. R., Pelling M. and Malamud B.D. (2009). Associations between elevated atmospheric temperature and human mortality: a critical review of the literature. *Climatic Change*, **92**, 299–341.

Hajat, S., Vardoulakis, S., Heaviside, C., Eggen, B. (2014). Climate change effects on human health: Projections of temperature-related mortality for the UK during the 2020s, 2050s and 2080s. *Journal of Epidemiology and Community Health*, **68**, 641-648.

Heaviside C., Tsangari H., Paschalidou A.K., Vardoulakis S., Kassomenos A.P., Georgiou K.E., Yamasaki E.N. (2016). Heat related mortality in Cyprus for current and future climate scenarios. *Science of the Total Environment*, **569-570**, 627-633.

Iñiguez, C., Ballester, F., Ferrandiz, J., Pérez-Hoyos, S., Sáez, M., López, A. (2010). Relation between Temperature and Mortality in Thirteen Spanish Cities. *International Journal of Environmental Research and Public Health*, **7,** 3196-3210.

IPCC (2013). Summary for Policymakers. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Kovats R. S. and Hajat S. (2008). Heat Stress and Public Health: A Critical Review. *Annual Review of Public Health*, **29**, 41-55.

Tsangari, H., Paschalidou, A., Vardoulakis, S., Heaviside, C., Konsoula, Z., Christou, S., Georgiou, K.E., Ioannou, K., Mesimeris, T., Kleanthous, S., Pashiardis, S., Pavlou, P., Kassomenos, P., Yamasaki, E.N. (2016). Human mortality in Cyprus - the role of temperature and particulate air pollution. *Regional Environmental Change*, **16**, 1905-1913.

Zachariadis, T. (2012). Climate change in Cyprus: Impacts and Adaptation Policies. *Cyprus Economic Policy Review*, **6**, 21-37.

Zanobetti A. and Schwartz, J. (2008). Temperature and mortality in nine US cities. *Epidemiology*, **19**, 563-570.

ΕΥΡΕΤΗΡΙΟ ΣΥΓΓΡΑΦΕΩΝ
(AUTHOR INDEX)