ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ & ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών στη Διοίκηση και τον Χρηματοοικονομικό Σχεδιασμό για στελέχη του Δημοσίου και Ιδιωτικού Τομέα

Executive MBA in Financial Planning

Μεταπτυχιακή Διατριβή

Πρόβλεψη αποτελέσματος σε ποδοσφαιρικούς αγώνες με τη χρήση μοντέλων παλινδρόμησης

Γεωργακόπουλος Σωτήριος

Γιακουμάτος Στέφανος: Γεωργακόπουλος Σωτήριος

Διατριβή υποβληθείσα στο Τμήμα Λογιστικής & Χρηματοοικονομικής του Πανεπιστημίου Πελοποννήσου. Η παρούσα διατριβή αποτελεί μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος στη Διοίκηση και τον Χρηματοοικονομικό Σχεδιασμό για στελέχη του Δημοσίου και Ιδιωτικού Τομέα

Καλαμάτα, Οκτώβριος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ & ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών στη Διοίκηση και τον Χρηματοοικονομικό Σχεδιασμό για στελέχη του Δημοσίου και Ιδιωτικού Τομέα

Executive MBA in Financial Planning

Τριμελής Εξεταστική Επιτροπή

Γιακουμάτος Στέφανος (Επιβλέπων)
Καθηγητής, Τμήμα Λογιστικής Χρηματοοικονομικής, Πανεπιστήμιο Πελοποννήσου

Γιαννόπουλος Βασίλειος
Επίκουρος καθηγητής, Τμήμα Λογιστικής Χρηματοοικονομικής, Πανεπιστήμιο Πελοποννήσου

Μαυριδόγλου Γεώργιος
Λέκτορας, Τμήμα Λογιστικής Χρηματοοικονομικής, Πανεπιστήμιο Πελοποννήσου

Ο Γεωργακόπουλος Σωτήριος

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι το Τμήμα Λογιστικής & Χρηματοοικονομικής μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη του Ιδρύματος, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Γιακουμάτο Στέφανο για την αμέριστη βοήθειά του, την άψογη καθοδήγησή του και τη στήριξή του κατά τη διάρκεια συγγραφής της παρούσας διπλωματικής εργασίας.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη στα Ελληνικά

Η παρούσα διπλωματική εργασία πραγματεύεται την επεξεργασία δεδομένων από αθλητικούς αγώνες ποδοσφαίρου. Τα δεδομένα υποβάλλονται σε στατιστική επεξεργασία, μέσω κατάλληλων στατιστικών μεθόδων. Τα μοντέλα που χρησιμοποιούνται σε όλη την έκταση της εργασίας είναι μοντέλα παλινδρόμησης, ενώ δίνεται έμφαση στις κατανομές Poisson και αρνητική διωνυμική. Τα μοντέλα παλινδρόμησης μοντελοποιούν την τιμή μιας μεταβλητής η οποία εξαρτάται από ένα σύνολο άλλων μεταβλητών, ενώ στοχεύουν να βρουν με ακρίβεια την τάση των δεδομένων και να κάνουν προβλέψεις από παλαιότερα δεδομένα.

Τα δεδομένα που αναλύονται αφορούν το τελικό αποτέλεσμα αγώνων ποδοσφαίρου, δηλαδή πόσα γκολ σημείωσαν οι αντίπαλες ομάδες. Τα αποτελέσματα της μοντελοποίησης παρουσιάζονται μέσα από γραφήματα και κατάλληλα λογισμικά στατιστικά πακέτα. Από την στατιστική μοντελοποίηση των δεδομένων και την ερμηνεία των αποτελεσμάτων επιδιώκεται η πρόβλεψη του τελικού αποτελέσματος σε ένα αγώνα ποδοσφαίρου. Παράλληλα αναδεικνύεται μια ενιαία θεωρητική και εννοιολογική δομή για μοντελοποίηση στη στατιστική, ώστε ο αναγνώστης να κατανοήσει περισσότερο τις στατιστικές τεχνικές που χρησιμοποιούνται στην ανάλυση δεδομένων και να αποκομίσει τις περισσότερες δυνατές πληροφορίες.

Λέξεις κλειδιά: Μοντέλα, Παλινδρόμηση, Κατανομή, Πρόβλεψη, Ποδόσφαιρο

Abstract

This master thesis deals with athlete data processing of football matches. The data shall be submitted to processing through appropriate statistical methods. The models used in this thesis are regression models and the accent is given on Poisson and negative binomial distributions. Regression models model the value of a variable, which depends on other variables, while the models target is to find accurately the data trend and make predictions from older data.

The data analyzed in the paper are related to the final football score, that is how many goals scored by the opponents teams. The results of modeling are presented through graphs and appropriate statistical software packages. The aim is to provide α prediction for the final result of a football match from the statistical data modeling and the interpretation of results. Furthermore a unified theoretical structure is emerging, for modeling in statistics, so the reader understands more the statistical techniques that been used in data analysis and derive as much information as possible.

Keywords: Models, Regression, Distribution, Predict, Football

ΚΑΤΑΛΟΓΟΣ ΓΡΑΦΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Εισαγωγή

Από στατιστική άποψη η μοντελοποίηση είναι μια μέθοδος της ανάλυσης δεδομένων. Το ζητούμενο στην μοντελοποίση είναι να βρεθεί ένα μαθηματικό υπόδειγμα το οποίο ταιριάζει με τα δεδομένα και μπορεί να τα περιγράψει με τον καλύτερο δυνατό τρόπο. Μετέπειτα γίνεται εκτίμηση των παραμέτρων του μοντέλου, και στην συνέχεια το μοντέλο έρχεται αντιμέτωπο με τα δεδομένα μέσω του επονομαζόμενου ελέγχου, ο οποίος έρχεται να δικαιολογήσει το μοντέλο. Η επιτυχία σε μια τέτοια προσπάθεια έγκειται στην επιλογή και την κατασκευή ενός υποδείγματος το οποίο περιλαμβάνει όσο το δυνατόν περισσότερη πληροφορία σε σχέση με την πραγματικότητα, όπως αυτή αποτυπώνεται με βάση τα διαθέσιμα δεδομένα. Πρόκειται για ένα σύστημα μαθηματικών σχέσεων και εξισώσεων που αποσκοπούν στην προσομοίωση του αντίστοιχου φαινομένου.

Η μοντελοποίηση του αριθμού των γκολ που πετυχαίνει μια ομάδα είναι ένα δημοφιλές θέμα και πολλές μελέτες έχουν γίνει για το σκοπό αυτό. Στην μοντελοποίηση του αριθμού των γκολ μια θεμελιώδης ερώτηση είναι ποια κατανομή μπορεί να χρησιμοποιηθεί, η οποία ταιριάζει με την πραγματική κατανομή των γκολ. Τα περισσότερα μοντέλα που υπάρχουν στην βιβλιογραφία για εκτίμηση των γκολ βασίζονται στην κατανομή Poisson. Η Poisson κατανομή έχει μια επίσημη θεωρητική βάση και χρησιμοποιείται για γεγονότα που συμβαίνουν τυχαία και σε ένα σταθερό ρυθμό, σε παρατηρούμενη χρονική περίοδο. Αυτό είναι ισοδύναμο με το να υποθέσουμε ότι η ικανότητα σκοραρίσματος της ομάδας είναι σταθερή σε όλη τη σεζόν. Αυτή η υπόθεση είναι περιοριστική διότι η δύναμη κάθε ομάδας αλλάζει από παιχνίδι σε παιχνίδι. Η υπόθεση της μεταβολής της ικανότητας της ομάδας οδηγεί σε άλλες κατανομές, με πιο κατάλληλη κατανομή την αρνητική διωνυμική ως εναλλακτική της Poisson κατανομής. Ωστόσο αποδεικνύεται ότι οι δύο κατανομές δεν παρουσιάζουν μεγάλες διαφορές όσον αφορά τις τελικές προβλέψεις. Επιπλέον η Poisson προσφέρει σημαντικά μεγαλύτερη ευκολία στους υπολογισμούς σε σχέση με την αρνητική διωνυμική. Για τους παραπάνω λόγους η κατανομή Poisson έχει γίνει ευρέως αποδεκτή ως κατάλληλο μοντέλο για τα γκολ σε ένα αγώνα ποδοσφαίρου. Οι δύο κατανομές χρησιμοποιούνται σε πολλές περιπτώσεις, όταν εξετάζονται δεδομένα καταμέτρησης, ενώ αποτελούν σημαντικές κατανομές στη στατιστική συμπερασματολογία.

Στο πρώτο κεφάλαιο παρουσιάζονται οι μέθοδοι ανάλυσης δεδομένων. Στο δεύτερο κεφάλαιο παρουσιάζεται το μοντέλο του Maher (1982), το οποίο αποτελεί την πρώτη προσπάθεια για μοντελοποίηση στο ποδόσφαιρο. Το τρίτο κεφάλαιο παρέχει το θεωρητικό πλάισιο των στατιστικών τεχνικών στην μοντελοποίηση, ενώ στο τέταρτο κεφάλαιο αναλύονται

τα γενικευμένα γραμμικά μοντέλα. Τέλος στο πέμπτο κεφάλαιο γίνεται εφαρμογή του μοντέλου του Maher σε πραγματικά δεδομένα από το Ελληνικό πρωτάθλημα.

Κεφάλαιο 1 Ανάλυση δεδομένων και στατιστική

1.1 Στατιστική

Στατιστική είναι η επιστήμη που προσπαθεί να εξαγάγει ασφαλή συμπεράσματα χρησιμοποιώντας εμπειρικά δεδομένα παρατήρησης και πειράματα. Η στατιστική μπορεί να ταξινομηθεί σε δύο τύπους, την περιγραφική στατιστκή και την επαγωγική στατιστική ή στατιστική συμπερασματολογία. Η περιγραφική στατιστική συλλέγει, οργανώνει, παρουσιάζει και περιγράφει τα δεδομένα μέσα από πίνακες, διαγράμματα και αριθμούς, όπως είναι η μέση τιμή, η διάμεσος και η διακύμανση. Δίνει πληροφορίες σχετικά με την ποιότητα των δεδομένων και απαντάει στην ερώτηση τι συνέβη. Ωστόσο δεν παρέχει τον λόγο γιατί κάτι συμβαίνει. Αυτό είναι κομμάτι της επαγωγικής στατιστικής ή στατιστικής συμπερασματολογίας, η οποία προσπαθεί να κάνει πρόβλεψη για τον πληθυσμό από τη μελέτη ενός μικρού δείγματος του πληθυσμού. Η συμπερασματολογία αναλύει δεδομένα και ενσωματώνει τάσεις και πρότυπα για να προβλέψει τι είναι πιθανό να συμβεί στο μέλλον, ενώ απαιτεί πιο πολύπλοκα εργαλεία για την ανάπτυξή της σε σύγκριση με την περιγραφική στατιστική (Κικίλιας et al, 2001).

1.2 Μέθοδοι ανάλυσης δεδομένων

Η ανάλυση δεδομένων είναι συνώνυμη με την επαγωγική στατιστική. Ωστόσο για ορισμένους η ανάλυση δεδομένων εκφράζει μια διαφορετική θεώρηση στη στατιστική συμπεραματολογία, εξαιτίας μεθόδων οι οποίες έρχονται σε αντίθεση με τον παραδοσιακό τρόπο των στατιστικών ελέγχων. Η ανάλυση δεδομένων έχει εξελιχθεί πολύ τα τελευταία χρόνια εξαιτίας των δυνατοτήτων που προσφέρουν οι ηλεκρονικοί υπολογιστές. Πολύπλοκοι αριθμητικοί υπολογισμοί μπορούν να γίνουν με την χρήση των υπολογιστών. Σύμφωνα με τον Μπεχράκη (1999) η ανάπτυξη και η γενικευμένη χρήση των ηλεκτρονικών υπολογιστών είναι ένας σημαντικός λόγος που η ανάλυση δεδομένων έχει υιοθετηθεί και εφαρμόζεται από πολλά διαφορετικά επιστημονικά πεδία. Τα τελευταία χρόνια έχουν αναπτυχθεί κατάλληλα προγράμματα, γνωστά ως στατιστικά πακέτα, όπως το SPPS, Statistica, SAS, STATA.

Μέθοδοι ανάλυσης δεδομένων, οι οποίες αναπτύχθηκαν μετά την δεκαετία του 1970, έχουν ως κύριο χαρακτηριστικό τους ότι δεν απαιτούν εκ των προτέρων την ύπαρξη κάποιας θεωρητικής κατανομής, ενώ δεν κάνουν διάκριση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών. Η ανάγκη να μην θεωρείται εκ των προτέρων ότι ένα φαινόμενο ακολουθεί κάποιο συγκεκριμένο νόμο, οδήγησε σε μεθόδους της ανάλυσης δεδομένων οι οποίες λέγονται μη παραμετρικές μέθοδοι ή στατιστική δίχως μοντέλα (Καραπιστόλης, 1999). Σύμφωνα με τον Καραπιστόλη η ανάλυση δεδομένων αποτελεί κλάδο της πολυδιάστατης στατιστικής ανάλυσης και περιλαμβάνει τις μεθόδους, της παραγοντικής ανάλυσης των αντιστοιχιών, την ανάλυση σε κύριες συστάδες και την ταξιμόμηση σε αύξουσα ιεραρχία. Σύμφωνα με τον Παπαδημητρίου

(2004), η παραγοντική ανάλυση αντιστοιχιών αποτελεί μια από τις σημαντικότερες μέθοδος με την οποία μπορεί να αναλυθεί ένα πολυμεταβλητό φαινόμενο. Σύμφωνα με τον Clausen (1998) η μέθοδος της παραγοντικής ανάλυσης αντιστοιχιών εμφανίστηκε και αναπτύχθηκε ανεξάρτητα και σε ορισμένες περιπτώσεις σχεδόν ταυτόχρονα σε πολλές χώρες όπως οι Η.Π.Α., η Μεγάλη Βρετανία, η Γαλλία, η Ιαπωνία, ο Καναδάς, η Ολλανδία.

Σκοπός των μεθόδων αυτών είναι να περιγράψουν πολυδιάστατους πίνακες δεδομένων μέσα από διαδικασίες ελάττωσης των διαστάσεων του αρχικού χώρου στον οποίο το υπό εξέταση φαινόμενο μπορεί να περιγραφεί. Σύμφωνα με τους Dillon και Goldstein (1984) οι νέες διαστάσεις οι οποίες προκύπτουν από πολύπλοκες σχέσεις μεταξύ των μετρήσεων ορίζονται τελικά ως νέες μεταβλητές ή παράγοντες. Σύμφωνα με τον Benzecri (1991) η διαφορετική αυτή θεώρηση δεν απαιτεί κάποια υπόθεση σχετικά με τις παραμέτρους του υπό εξέταση πληθυσμού, δηλαδή την ύπαρξη κάποιου στοχαστικού υποδείγματος. Επίσης δεν χρειάζεται έλεγχος υποθέσων όπως συνηθίζεται σε γνωστά μοντέλα της κλασικής στατιστικής που βασίζονται σε πιθανοθεωρητικές προσεγγίσεις. Σύμφωνα με τον Αθανασιάδη (1995) το χαρακτηριστικό των μεθόδων που αναπτύχθηκαν ήταν η προσήλωση στα δεδομένα, ενώ δεν δίνεται μεγάλη σημασία στο θεωρητικό πλαίσιο που οδήγησε στην συγκρότησή τους.

1.3 Η Γαλλική και η Ολλανδική σχολή ανάλυσης δεδομένων

Σύμφωνα με τον Γάλλο Benzecri (1973) η στατιστική και η θεωρία πιθανοτήτων δεν είναι το ίδιο, και η στατιστική βρίθει από υποθέσεις που ποτέ δεν ικανοποιούνται στην πράξη. Επίσης οι μέθοδοι που δεν περιλαμβάνουν πιθανοθεωρητικούς μηχανισμούς απελευθερώνουν τον ερευνητή από δεσμεύσεις και υποθέσεις που έχουν τα θεωρητικά μοντέλα, ενώ ο ίδιος ο χρήστης της μεθόδου αναλαμβάνει την ευθύνη να εξετάσει και να ερμηνεύσει τα δεδομένα. Ο Benzecri ισχυρίζεται ότι οι μεγάλες αδόμητες πολυμεταβλητές ομάδες δεδομένων είναι η μόνη δυνατή κατάσταση έρευνας, και ότι η στατιστική ανάλυση ιδίως όταν εμπλέκονται πολλές διαστάσεις θεμελιώνεται περισσότερο σε αλγεβρικούς και γεωμετρικούς υπολογισμούς παρά σε πιθανολογικούς.

Ερευνητές της Ολλανδικής σχολής ανάλυσης δεδομένων δεν συμφωνούν απόλυτα με τις θέσεις του Benzecri. Σύμφωνα με τους Ολλανδούς, αν οι νέες μέθοδοι εφαρμοστούν συστηματικά, και διαχωριστούν από τις κλασικές μεθόδους, είναι δυνατόν να οδηγήσουν σε ένα τυφλό εμπειρισμό και σε αυθαιρεσία. Επίσης κινδυνεύει να χαθεί η αυστηρή μαθηματική διατύπωση και η γενίκευση που υπάρχει στα θεωρητικά μοντέλα. Αντίθετα πιστεύουν ότι κάποιο θεωρητικό μοντέλο θα πρέπει να καθοδηγεί τον χρήστη των μεθόδων, ώστε να είναι δυνατή η ερμηνεία των αποτελεσμάτων, ενώ οι πιθανολογικές έννοιες και ιδέες είναι απαραίτητες διότι μπορούν να αναδείξουν την αναγκαιότητα των αλγεβρικών πράξεων και μερικές φορές μπορούν

να χρησιμοποιηθούν για να αξιολογήσουν την χρησιμότητά τους. Για τους Ολλανδούς ερευνητές η θεωρία πιθανοτήτων, η στατιστική, η υπολογιστική στατιστική, τα μαθηματικά, τα υπολογιστικά μαθηματικά και η πληροφορική, όχι μόνο ενσωματώνονται στην ανάλυση δεδομένων αλλά μπορούν να συνδυαστούν ώστε τα αποτελέσματα που παράγονται να είναι σταθερά και αξιόπιστα. Ο έλεγχος της σταθερότητας είναι σημαντικός στην ανάλυση δεδομένων και σημαίνει ότι τα αποτελέσματα πρέπει να συμφωνούν με τις αρχικές υποθέσεις και τους περιορισμούς. Τα αποτελέσματα πρέπει να είναι σταθερά στις περιπτώσεις όπου η έρευνα επαναλαμβάνεται, κάτω από τις ίδιες συνθήκες, με διαφορετική μέθοδο συλλογής δεδομένων, με διαφορετικά μοντέλα, με διαφορετική μέθοδο ανάλυσης δεδομένων, κάτω από την ισχύ των κεντρικών θεωρημάτων, κάτω από την επίδραση διαφορετικών μεθόδων ανάλυσης (Μενεξές, 2007).

1.4 Ανάλυση δεδομένων

Η ανάλυση δεδομένων σχετίζεται με την προσπάθεια κατανόησης της τάσης των δεδομένων. Για το σκοπό αυτό χρησιμοποιούνται θεωρητικά μαθηματικά μοντέλα που προσαρμόζονται στα δεδομένα. Με τη βοήθεια απλών συναρτήσεων στατιστικής ανάλυσης ή κατανομών, ως και πιο σύνθετες μεθόδους ανάλυσης και εξόρυξης δεδομένων, μπορεί να γίνει ανάλυση και περιγραφή της πολυπλοκότητας ορισμένων φαινομένων, προκειμένου να αναδειχθεί η συμπεριφορά τους. Οι παραδοσιακοί μέθοδοι βασίζονται στην επαγωγική στατιστική και σε μαθηματικά υποδείγματα που χρησιμοποιούν εκ των προτέρων υποθέσεις, έχουν πιθανολογικές προϋποθέσεις και εφαρμόζονται για την απόρριψη ή όχι συγκεκριμένων ερευνητικών υποθέσεων. Με αυτόν τον τρόπο υπάρχει εξ αρχής το μαθηματικό υπόδειγμα που προσαρμόζεται στα δεδομένα και μετά η συμπεραματολογία υπάγεται σε κάποιο μηχανισμό επαγωγικού συλλογισμού με την έννοια της στατιστικής σημαντικότητας μέσω στατιστικών ελέγχων που είναι απαιτήσεις των στατιστικών τεχνικών. Η αρχική θεώρηση του μοντέλου εξασφαλίζει ότι τα μοντέλα έχουν ελεχθεί ως προς τις ιδιότητες, τα κριτήριά τους και τον τρόπο που αξιολογούνται. Αυτόματα εισάγονται οι αρχικές υποθέσεις που εξασφαλίζουν την εγκυρότητα των μεθόδων ανάλυσης και αφορούν, τον τρόπο που θα αξιολογηθεί η πληροφορία, την αντιπροσωπευτικότητα των δεδομένων, την κωδικοποίησή τους, την δυνατότητα πραγματοποίησης υπολογισμών και την ακρίβειά τους (Μενεξές, 2007).

Οι χρήστες των μεθόδων μπορούν να επηράσουν τα αποτελέσματα της έρευνας ανάλογα με τις μεθόδους που θα επιλέξουν, και άρα θα πρέπει να αξιολογούν τις μεθόδους καθώς και τα αποτελέσματα. Για αυτό είναι σημαντική η θέσπιση κάποιων κανόνων και κριτηρίων. Αν υπάρχει κάποιο μοτίβο στα δεδομένα οι μέθοδοι θα το αναδείξουν. Ιδιαίτερη έμφαση δίνεται στις μεταβλητές που θα χρησιμοποιηθούν και θα περιγράψουν το υπό εξέταση φαινόμενο. Ίσως

χρειαστεί να εισαχθούν νέες μεταβλητές ή αντίθετα να αφαιρεθούν μεταβλητές που λειτουργούν συσκοτιστικά.

Ωστόσο δεν υπάρχει πάντα, και για οποιοδήποτε φαινόμενο, το αντίστοιχο μοντέλο. Οι μέθοδοι της ανάλυσης δεδομένων δεν διαθέτουν κάποιο μαγικό μηχανισμό, όπου κάθε φορά ένα αδόμητο σύνολο δεδομένων θα μεταραπεί σε επιστημονική γνώση. Σύμφωνα με τον Δερμάνη (1986) δεν γνωρίζουμε τους νόμους στους οποίους ένα σύνολο δεδομένων ή ένα φαινόμενο που εξετάζεται υπακούει σε ένα μαθηματικό υπόδειγμα ντετερμινιστικό ή στοχαστικό το οποίο καθορίζει τη συμπεριφορά του. Συνήθως τα δεδομένα είναι σύνθετα και δεν ταιριάζουν σε μια μαθηματική περιγραφή ή δεν μπορεί να διατυπωθεί κάποιο θεώρημα που να εκφράζει τη σχέση μεταξύ αιτίας και αποτελέσματος. Ακόμα και αν βρεθεί ένα μοντέλο που βοηθάει στην ανακάλυψη κάποιων σημαντικών στοιχείων, δεν αποδεικνύει τα πάντα για το πραγματικό φαινόμενο (De Leeuw, 2005). Σύμφωνα με τον Pfeiffer (1978) ένα υπόδειγμα δεν πρέπει να χαρακτηρίζεται ως σωστό ή λάθος, αλλά ως χρήσιμο ή μη χρήσιμο. Σύμφωνα με τον Breiman (2001) ιδιαίτερη έμφαση θα πρέπει να δίνεται στην κατανόηση του εκάστοτε προβλήματος και των αντίστοιχων εμπειρικών δεδομένων πριν την προσαρμογή και τον έλεγχο οποιουδήποτε υποδείγματος. Οι μέθοδοι ανάλυσης δεδομένων μπορούν να συμβάλλουν προς την κατεύθυνση αυτή αποκαλύπτοντας πληροφορία σχετικά με τον συχνά άγνωστο μηχανισμό παραγωγής δεδομένων.

Κεφάλαιο 2 Μοντελοποίηση στο ποδόσφαιρο

2.1 Κίνητρο για πρόβλεψη

2.1.1 Πρόβλεψη στις επιχειρήσεις

Οι άνθρωποι πάντα έψαχναν τα δεδομένα για να βρουν σχέσεις και να προβλέψουν μελλοντικά αποτελέσματα. Σήμερα η πρόοδος στην ανάλυση δεδομένων και στην ταχύτητα επέτρεψαν να βρεθούν καινούργιες μεθόδοι, σε αντίθεση με τους παλαιούς χρονοβόρους τρόπους. Η αυξανόμενη ανάγκη για πρόβλεψη στο μάρκετινγκ, σε τομείς υγείας, στην ασφάλεια, είχε ως αποτέλεσμα την ανάπτυξη μοντέλων πρόβλεψης δεδομένων.

Οι επιχειρήσεις χρησιμοποιούν στατιστική για να μελετήσουν παλαιά και νέα δεδομένα, για να κατανοήσουν την τάση τους, προκειμένου να αναπτύξουν επιχειρηματικό σχεδιασμό και να βελτιώσουν την απόδοσή τους. Η προγνωστική ανάλυση μελετάει παλαιά δεδομένα και ενσωματώνει τάσεις και πρότυπα για να προβλέψει τι είναι πιθανό να συμβεί στο μέλλον. Όπως οι επιχειρήσεις, έτσι και στα επαγγελματικά αθλήματα η χρήση της στατιστικής έχει γίνει όλο και πιο σημαντική. Ενώ πολλά δεδομένα ανάλυσης που χρησιμοποιούν οι επιχειρήσεις εμπίπτουν στην περιγραφική ανάλυση, οι περισσότερες έρευνες στα σπορ επικεντρώνονται σε μεθόδους πρόβλεψης, διότι η πρόβλεψη μελλοντικών αποδόσεων και οι μελλοντικές κινήσεις είναι κρίσιμες για να κερδίσει μια ομάδα.

2.1.2 Πρόβλεψη σε χαρακτηριστικά παιχνιδιού

Σήμερα εταιρίες έχουν δημιουργήσει τεράστιες βάσεις δεδομένων με πληροφορίες από διαφορετικά γεγονότα που εμφανίζονται σε αγώνες ποδοσφαίρου. Είναι εύκολο να βρει κάποιος πληροφορίες που έχουν να κάνουν με τα αποτελέσματα των αγώνων, τις πάσες, την κατοχή της μπάλας, την συμπεριφορά του παιχνιδιού, τις τακτικές του προπονητή. Ιδίως τα τελευταία χρόνια είναι αξιοσημείωτος ο τρόπος που συλλέγονται υψηλής ποιότητας δεδομένα από παιχνίδια, χάρη σε αυτοματοποιημένες αισθητηριακές τεχνικές που βασίζονται σε καταγραφή βίντεο ή ακόμα και σε παρατηρήσεις από ποικίλους τρόπους σταθερών και κινητών αισθητήρων. Σκοπός της ανάλυσης αθλητικών δεδομένων είναι η μέτρηση της απόδοσης της ομάδας και κάθε παίχτη ατομικά, η πρόβλεψη νίκης ή ήττας, η πρόβλεψη των επόμενων κινήσεων της αντίπαλης ομάδας, η αποτροπή τραυματισμών, η βελτίωση της προπόνησης καθώς και οτιδήποτε άλλο θα μπορούσε να βοηθήσει την ομάδα αυξάνοντας την πιθανότητα νίκης.

Η χρήση της ανάλυση αθλητικών δεδομένων έχει αλλάξει τον τρόπο που παίζεται κάθε παιχνίδι, διότι οι αποφάσεις σε όλες τις πτυχές του παιχνιδιού βασίζονται τα τελευταία χρόνια σε στατιστικά δεδομένα και προβλέψεις. Η ανάλυση δεδομένων μπορεί να βοηθήσει μια ομάδα,

τόσο εντός όσο και εκτός γηπέδου, προκειμένου να αποκτήσει γνώσεις και σε πολλές περιπτώσεις ανταγωνιστικό πλεονέκτημα. Η απόφαση για την καλύτερη δυνατή σύνθεση των ομάδων, την καλύτερη απόδοση των παιχτών ή την καλύτερη δυνατή στρατηγική της ομάδας γίνεται πλέον όχι μόνο από την κρίση του προπονητή αλλά από μετρήσεις που βασίζονται σε δεξιότητες.

Ωστόσο, ανεξάρτητα από την τεράστια ποσότητα δεδομένων, το ένστικτο πολλές φορές δεν μπορεί να αντικατασταθεί από στατιστικά. Σύμφωνα με τον Davenport (2014) οι αλγόριθμοι και οι υπολογιστές ενδέχεται να αγνοήσουν μεταβλητές όταν προβλέπουν μια αθλητική συμπεριφορά. Επίσης η απόδοση ενός παίχτη δεν μπορεί να εκτιμηθεί ακριβώς και μόνο με μετρήσεις και με απόλυτους αριθμούς. Υπάρχουν παίχτες που λειτουργούν εμπειρικά και από ένστικτο και, ενώ δεν έχουν τα καλύτερα στατιστικά, μπορεί από μια ενέργειά τους να κρίνουν το αποτέλεσμα. Παρ'όλα αυτά η ανάλυση δεδομένων μπορεί να λειτουργήσει συμπληρωματικά ως ένας σημαντικός παράγοντας ενίσχυσης για την επιτυχία.

2.1.3 Πρόβλεψη και λόγοι στοιχηματισμού

Το στοίχημα σε αθλήματα έχει γίνει πολύ δημοφιλές τα τελευταία χρόνια, με πολλούς να παίζουν τακτικά ή περιστασιακά. Εταιρίες στοιχημάτων προσφέρουν αποδόσεις και παίρνουν στοιχήματα για πολλές αθλητικές εκδηλώσεις, ενώ εστιάζουν κυρίως σε στοιχήματα από τον επαγγελματικό αθλητισμό και το ποδόσφαιρο. Πράκτορες προσφέρουν πιθανότητες σε πολλά αποτελέσματα ενός αγώνα. Η πιο απλή εκδοχή χρησιμοποιεί το αποτέλεσμα του αγώνα, νίκη, ήττα ή ισοπαλία, αλλά μπορεί να γίνει στοιχηματισμός και σε πολλά άλλα χαρακτηριστικά όπως το σκορ, ο αριθμός των κίτρινων καρτών και των κόρνερ. Οι προβλέψεις αγώνων ποδοσφαίρου έχουν αποκτήσει τεράστιο ενδιαφέρον για τους φιλάθλους και για την βιομηχανία στοιχημάτων (Καρλής & Ντζούφρας, 2008).

2.1.4 Πρόβλεψη στο ποδόσφαιρο

Το ποδόσφαιρο είναι το πιο δημοφιλές άθλημα στον κόσμο. Μια τεράστια αγορά έχει δημιουργηθεί γύρω από το ποδόσφαιρο, ενώ οι αθλητές αμείβονται με αστρονομικά ποσά. Εκτός από το γεγονός ότι είναι ένα συναρπαστικό άθλημα, έχει και ένα άλλο στοιχείο που το ξεχωρίζει και κινεί το ενδιαφέρον όλων. Το χαρακτηριστικό του ποδοσφαίρου είναι ότι το κύριο γεγονός του, δηλαδή το γκολ, είναι σπάνιο. Τα συνηθισμένα τελικά αποτελέσματα σε ένα αγώνα είναι 0-0, 1-0, 0-1, 1-1 και αυτά κρίνουν το ποιος θα είναι ο καλύτερος. Άρα τα σημαντικά γεγονότα είναι λίγα σε σχέση με άλλα αθλήματα, όπως για παράδειγμα το μπάσκετ όπου μπαίνουν πολλά καλάθια, δηλαδή υπάρχουν πολλά γεγονότα που μπορούν να καταγραφούν.

Επειδή τα γεγονότα είναι λίγα, είναι δύσκολο να γίνει πρόβλεψη και υπάρχει μεγάλη αβεβαιότητα. Ακόμα και ο αδύναμος έχει πιθανότητες και ελπίδα να κερδίσει. Αυτός είναι ένας σημαντικός λόγος που το ποδόσφαιρο είναι τόσο δημοφιλές, εκτός από την ίδια του την ομορφιά σαν άθλημα. Η πρόβλεψη στο ποδόσφαιρο είναι κάτι που απασχολεί πολλούς, αν και είναι πολύ δύσκολο να γίνει. Πολλοί στατιστικοί ιδίως τα τελευταία χρόνια αναλύουν δεδομένα ποδοσφαίρου με σκοπό να βρουν μοντέλα που προβλέπουν το αποτέλεσματα σε αγώνες ποδοσφαίρου (Ντζούφρας, 2019).

2.2 Στατιστικά μοντέλα για ποδόσφαιρο

2.2.1 Μοντέλα για τα χαρακτηριστικά του παιχνιδιού

Υπάρχουν πολλές στατιστικές τεχνικές στην μοντελοποίηση δεδομένων ποδοσφαίρου. Οι πιο πολλές αφορούν τον αριθμού των γκολ που μπαίνουν σε κάθε παιχνίδι. Ωστόσο υπάρχουν στατιστικά μοντέλα για πολλά χαρακτηριστικά του παιχνιδιού. Για παράδειγμα μεγάλη σημασία έχουν τα μοντέλα για την αξία του παίχτη. Μια ομάδα πρέπει να ξέρει αν θα πουλήσει ή θα αγοράσει ένα παίχτη, αλλά και το πότε θα τον πουλήσει. Επίσης πρέπει να κάνει καλό σκάουτινγκ, δηλαδή να βρει παίχτες που θα τους αγοράσει φθηνά και θα τους πουλήσει ακριβά. Η αξιολόγηση των παιχτών τα τελευταία χρόνια γίνεται από δεδομένα που προέρχονται από μετρήσεις, και όχι μόνο διαισθητικά ή εμπειρικά. Ρούχα και ρολόγια φέρουν τσιπάκια που καταγράφουν τα πάντα στον αγωνιστικό χώρο, και τα αποτελέσματα αυτά χρησιμοποιούνται για την αξιολόγηση των παιχτών και την βελτίωση των δεξιοτήτων τους. Οι επιδόσεις των παιχτών μετριούνται και αξιολογούνται, έτσι ώστε ο προπονητής να μπορεί να πάρει τις σωστές αποφάσεις για το ποιος παίχτης αποδίδει καλύτερα. Το επαγγελματικό πρωτάθλημα μπάσκετ NBA στην Αμερική έχει τεράστια βιβλιογραφία όσον αφορά μοντέλα για την αξιολόγηση των παιχτών (Ντζούφρας, 2019).

Υπάρχει επίσης στατιστικό μοντέλο για το πρόγραμμα του πρωταθλήματος. Αν το πρόγραμμα δεν φτιαχτεί με τον σωστό τρόπο, κάποιες ομάδες μπορεί να καταλήξουν να παίζουν δύσκολους αγώνες τον έναν μετά τον άλλο και έτσι το πρωτάθλημα να γίνει μεροληπτικό, ή μπορεί να κινδυνέυει να χάσει το ενδιαφέρον του γιατί όλοι οι καλοί αγώνες θα έχουν μεταφερθεί στην αρχή του πρωταθλήματος. Πολλά διαφορετικά μοντέλα περιλαμβάνουν, δίκτυα πασών, την επίδραση της κόκκινης κάρτας, την επίδραση του γηπέδου, την επίδραση διαιτητών, την επίδραση έδρας, ανάλυση για το αν οι αγώνες είναι στημένοι, ανάλυση για το χρόνο που θα γίνουν οι αλλαγές των παιχτών κατά τη διάρκεια ενός αγώνα.

2.2.2 Μοντέλο για τα γκολ και μοντέλο για το αποτέλεσμα αγώνα

Η μοντελοποίηση σε ένα αγώνα ποδοσφαίρου εστιάζει στον αριθμό των γκολ που βάζει και δέχεται μια ομάδα. Χρησιμοποιεί διμεταβλητή Poisson παλινδρόμηση για να εκτιμήσει τα γκολ, ενώ η πρόβλεψη νίκης, ισοπαλίας ή ήττας προκύπτει έμμεσα αθροίζοντας κατάλληλα τις εκτιμώμενες καθορισμένες πιθανότητες. Υπάρχει επίσης μια δεύτερη προσέγγιση που αναπτύχθηκε πρόσφατα και περιλαμβάνει απευθείας μοντελοποίηση της πιθανότητας νίκης, ισοπαλίας ή ήττας, χρησιμοποιώντας για εκτίμηση μοντέλο παλινδρόμησης Probit (Goddard, 2005). Τα δύο μοντέλα φτιάχτηκαν ανεξάρτητα και αποδείχθηκε ότι δίνουν παρόμοια αποτελέσματα. Ωστόσο η μοντελοποίηση των γκολ, και όχι απευθείας του αποτελέσματος, προτιμάται από τους στατιστικούς και έχει περισσότερη βιβλιογραφία. Η πρώτη απόπειρα για μοντελοποίηση των γκολ στο ποδόσφαιρο έγινε με την εργασία του Maher το 1982.

2.3 Μοντέλο για τον αριθμό των γκολ

2.3.1 Το μοντέλο του Maher

Σύμφωνα με τους Dixon και Coles (1997) ένα στατιστικό μοντέλο για αγώνες ποδοσφαίρου απαιτεί πολλά χαρακτηριστικά. Για παράδειγμα το μοντέλο πρέπει να μετράει τη δύναμη των δύο ομάδων σε ένα παιχνίδι. Η ικανότητα αυτή της ομάδας πρέπει να συνοψίζεται σε διαφορετικά μέτρα για την επίθεση και την άμυνα, δηλαδή για τα γκολ που βάζει και δέχεται αντίστοιχα, ενώ η ικανότητα αυτή πρέπει να προκύπτει από πρόσφατα αποτελέσματα και όχι πολύ παλαιά. Επίσης το μοντέλο πρέπει να μετράει ένα πλεονέκτημα για το γεγονός ότι μια ομάδα παίζει εντός έδρας. Εμπειρικά είναι δύσκολο να υπολογιστεί η πιθανότητα αποτελέσματος ενός αγώνα κάτω από όλους αυτούς τους περιορισμούς. Το ζητούμενο είναι να βρεθεί ένα μοντέλο που να μπορεί να ενσωματώσει αυτά τα χαρακτηριστικά.

Αυτός που έβαλε τις προδιαγραφές για μοντελοποίηση των γκολ ήταν ο Maher το 1982. Το αρχικό μοντέλο του Maher υποθέτει ότι τα γκολ ακολουθούν μια Poisson κατανομή όπου η μέση τιμή εξαρτάται από την δύναμη των ομάδων, δηλαδή από την αντίστοιχη επιθετική και αμυντική ικανότητα κάθε μέρους. Σύμφωνα με τον Maher (1982), << *Υπάρχουν καλοί λόγοι για να πιστέψουμε ότι ο αριθμός των γκολ μιας ομάδας είναι Poisson μεταβλητή. Η κατοχή είναι σημαντικός παράγοντας στο ποδόσφαιρο και κάθε φορά που η ομάδα έχει την μπάλα έχει πιθανότητα να επιτεθεί και να βάλει γκολ. Η πιθανότητα ρ ότι η επίθεση θα έχει αποτέλεσμα γκολ είναι φυσικά μικρή αλλά ο αριθμός των φορών που η ομάδα θα έχει την κατοχή στο παιχνίδι είναι πολύ μεγάλος. Αν ρ είναι σταθερό και οι επιθέσεις ανεξάρτητες ο αριθμός των γκολ θα είναι 0, 1, 2, 3... και τότε η Poisson προσέγγιση θα είναι καλή. Αν υποθέσουμε ότι η ικανότητα μιας ομάδας*

να βάζει γκολ δηλαδή η πιθανότητα σκοραρίσματος είναι σταθερή σε όλο το πρωτάθλημα, τότε η Poisson κατανομή είναι αληθοφανής >>.

Ωστόσο αυτή η υπόθεση είναι περιοριστική διότι, όπως οι περισσότεροι φίλαθλοι γνωρίζουν, οι καλύτερες ομάδες βάζουν πιο πολλά γκολ όταν παίζουν με αδύναμες ομάδες παρά με δυνατές. Επίσης οι ομάδες συνηθίζουν να βάζουν πιο πολλά γκολ όταν παίζουν εντός έδρας παρά εκτός έδρας. Μια κατανομή που θα μπορούσε να χρησιμοποιηθεί για να ξεπεράσει αυτό το πρόβλημα είναι η αρνητική διωνυμική. Σύμφωνα με τον Maher η μέση τιμή της Poisson θα μεταβάλλεται ανάλογα με την ποιότητα της ομάδας, οπότε αν πρέπει να επιλεγεί μια κατανομή των γκολ με μεταβαλλόμενη μέση τιμή, τότε πρέπει να είναι η διωνυμική. Ωστόστο ο Maher χρησιμοποίησε Poisson κατανομή και απέδειξε ότι δεν έχει μεγάλη απόκλιση από την αρνητική διωνυμική.

Το μοντέλο του Maher υποθέτει ότι οι δύο αριθμοί, γκολ εντός και γκολ εκτός έδρας, είναι τυχαίες ανεξάρτητες μεταβλητές που ακολουθούν την οριακή Poisson κατανομή. Το τελικό αποτέλεσμα του αγώνα τότε θεωρείται ότι ακολουθεί διμεταβλητή Poisson κατανομή. Αφού οι δύο τυχαίες μεταβλητές θεωρούνται ανεξάρτητες, η διμεταβλητή Poisson συνάρτητση πιθανότητας θα είναι απλά το αποτέλεσμα δύο οριακών Poisson πυκνοτήτων. Μαθηματικά αυτό σημαίνει:

$$X = \text{αριθμός γκολ εντός έδρας} \sim \text{Poisson}\ (\lambda_x)$$

$$Y = \text{αριθμός γκολ εκτός έδρας} \sim \text{Poisson}\ (\lambda_y)$$

$$P(X = x) = \frac{(\lambda_x))^x e^{-\lambda_x}}{x!}\ , \qquad P(Y = y) = \frac{(\lambda_y)^y e^{-\lambda_y}}{y!}$$

$$P(X=x, Y=y) = P(X=x)\cdot P(Y=y) = \frac{(\lambda_x)^x e^{-\lambda_x}}{x!} \cdot \frac{(\lambda_y)^y e^{-\lambda_y}}{y!}$$

Σύμφωνα με τον Maher (1982) αν η i ομάδα παίζει εντός έδρας εναντίον της j ομάδας που παίζει εκτός έδρας τότε:

$\lambda_x = \alpha_i \cdot \beta_j$

$\lambda_y = \gamma_i \cdot \delta_j$

όπου

α_i : η δύναμη την εντός ομάδας i στην επίθεση

β_j : η αδυναμία της εκτός ομάδας j στην άμυνα

γ_i : η αδυναμία της εντός ομάδας i στην άμυνα

δ_j : η δύναμη της εκτός ομάδας j στην επίθεση

Οι παράμετροι ικανοποιούν τους περιορισμούς $\sum_i \alpha_i = \sum_i \beta_i$ και $\sum_i \gamma_i = \sum_i \delta_i$, και επειδή τα X και Y έχουν θεωρηθεί ανεξάρτητα μεταξύ τους, διότι αντιπροσωπεύουν διαφορετικά παιχνίδια στα δύο τέρματα, η εκτίμηση των α και β θα είναι εξολοκλήρου από το x και η εκτίμηση των γ και δ από το y μόνο. Ο Maher εφάρμοσε ένα γενικευμένο γραμμικό μοντέλο και οι μέσες τιμές δίνονται από τις σχέσεις:

$\log(\lambda_x) = \alpha_i + \beta_j$ και

$\log(\lambda_y) = \delta_j + \gamma_i$

Όσο μεγαλύτερη είναι η δύναμη της επίθεσης τόσο μεγαλύτεροι είναι οι παράμετροι επίθεσης α_i και δ_j, ενώ όσο καλύτερη είναι η άμυνα τόσο μικρότεροι είναι οι αμυντικοί παράμετροι β_j και γ_i.

Για τον Maher είναι σημαντικό να προσθέσεις ένα σταθερό παράγοντα για όλες τις ομάδες που παίζουν εντός έδρας, οπότε η πρώτη εξίσωση στο γενικευμένο γραμμικό μοντέλο γίνεται $\log(\lambda_x) = \alpha_i + \beta_j + \kappa$, όπου κ η παράμετρος της επίδρασης έδρας. Ακολουθώντας αυτή την προσέγγιση, οι Saraiva et al (2016) παρουσίασαν μια μελέτη για τους αγώνες Ευρωπαικών πρωταθλημάτων. Τα αποτελέσματα έδειξαν ότι το 50% των αγώνων κερδίζουν ομάδες που παίζουν εντός έδρας, το 25% των αγώνων είναι ισοπαλία και 25% η εντός ομάδα έχασε. Άρα για κάποιο λόγο υπάρχει ένα εγγενές πλεονέκτημα για την ομάδα όταν παίζει εντός. Με αυτό τον τρόπο η επίδραση της έδρας μπορεί να εισαχθεί στο μοντέλο προκειμένου να προβλέψει τις πιθανότητες νίκης, ήττας, ισοπαλίας.

2.3.2 Δυναμικό μοντέλο

Οι δείκτες α, β, γ, δ αποδεικνύεται ότι αποτυπώνουν πολύ καλά την απόδοση της ομάδας, όμως δεν μεταβάλλονται με το χρόνο και, ως εκ τούτου, η απόδοση της ομάδας θεωρείται ότι είναι σταθερή κατά τη διάρκεια του πρωταθλήματος. Σύμφωνα με τους Baker and Mchale (2015) στην πραγματικότητα η δύναμη μιας ομάδας μεταβάλλεται σε κάθε παιχνίδι ανάλογα με την έδρα, την ικανότητα σκοραρίσματος και άμυνας και από άλλους παράγοντες όπως για παράδειγμα όταν οι ομάδες αγοράζουν και πουλάνε παίχτες, από τραυματισμούς, την καλή ή κακή φόρμα της ομάδας, την αποχώρηση του μάνατζερ της ομάδας. Η απόδοση της ομάδας τείνει να είναι δυναμική, μεταβάλλεται από μια περίοδο χρονική στην άλλη και αυτή η συμπεριφορά πρέπει να ενσωματώνεται στο μοντέλο. Σύμφωνα με τους Dixon και Coles (1997) η απόδοση της ομάδας είναι πιθανό να είναι πιο κοντά σχετιζόμενη με την απόδοσή της στους πρόσφατους αγώνες παρά στους προηγούμενους αγώνες και άρα η εκτίμηση των παραμέτρων πρέπει να γίνεται για κάθε χρονικό σημείο t που βασίζεται σε αποτελέσματα στο πρόσφατο παρελθόν πριν τον χρόνο t. Ωστόσο μοντέλα με δυναμικές παραμέτρους δεν είναι χρήσιμα για

πρόβλεψη διότι χρειάζονται όλα τα δεδομένα του πρωταθλήματος, τα οποία δεν είναι διαθέσιμα, προκειμένου να εκτιμηθούν οι παράμετροι. Για ευκολία θεωρείται ότι οι παράμετροι είναι σε χαλαρή αίσθηση τοπικά σταθεροί μέσα στο χρόνο και ότι η παλαιότερη πληροφορία έχει λιγότερη αξία από την πρόσφατη πληροφορία. Στα απλά μοντέλα οι παράμετροι της επίθεσης και άμυνας για κάθε ομάδα συνήθως εκτιμώνται εκ των υστέρων, μετά από ένα ολοκληρωμένο σετ δεδομένων για κάθε σεζόν που έχει συλλεχθεί.

2.3.3 Η νέα πρόταση με το διμεταβλητό Poisson μοντέλο

Ο Maher (1982) αξιολογώντας το ανεξάρτητο Poisson μοντέλο βρήκε μια μικρή έλλειψη προασαρμογής και συμπέρανε ότι η υπόθεση ανεξαρτησίας δεν είναι απολύτως σωστή. Όταν εξέτασε τις διαφορές των γκολ Z = X-Y παρατήρησε ότι το μοντέλο υποεκτιμά τις ισοπαλίες Z = X-Y = 0, δηλαδή οι πραγματικές ισοπαλίες είναι αισθητά περισσότερες από αυτές που εκτιμά το μοντέλο. Παρατήρησε ότι αυτό ίσως οφείλεται στο γεγονός ότι υπάρχει μια συσχέτιση μεταξύ των γκολ που μπαίνουν εντός και εκτός έδρας, δηλαδή μεταξύ X και Y. Για το λόγο αυτό πρότεινε ένα άλλο μοντέλο, το διμεταβλητό Poisson μοντέλο ως προέκταση του αρχικού. Το διμεταβλητό Poisson μοντέλο έχει οριακές κατανομές Poisson με μέσους μ και λ, αλλά υπάρχει μια συσχέτιση p μεταξύ του σκορ. Το ζεύγος X, Y μπορεί να γραφτεί $X = U+W$, $Y=V+W$, όπου U, V, W ανεξάρτητες Poisson με μέσους μ-n, λ-n, n αντίστοιχα και $n = p\cdot\sqrt{\mu\cdot\lambda}$ είναι η συνδιακύμανση μεταξύ X, Y (Maher, 1982). Τα αποτελέσματα της διμεταβλητής Poisson έδειξαν ότι η εισαγωγή της έξτρα παραμέτρου ρ βελτιώνει την εκτίμηση για τις συχνότητες και άρα βελτιώνει το ανεξάρτητο Poisson.

2.3.4 Μεταγενέστερες προσπάθειες

Ο Maher παρουσιάζοντας τη μέθοδο με ολοκληρωμένη σκοπιά, της έδωσε την αξία που της αρμόζει και παρότρυνε με τον τρόπο αυτό πολλούς ακόμα σημαντικούς στατιστικούς να την βελτιώσουν και να την αναπτύξουν. Μεταγενέστερες προσπάθειες ενίσχυσαν το μοντέλο του Maher σε μια ποικιλία κατευθύνσεων. Οι Dixon and Coles (1997) τροποποίησαν το μοντέλο του Maher. Βρήκαν ότι υπάρχει παρέκκλιση της υπόθεσης ανεξαρτησίας μεταξύ των γκολ που βάζουν οι δύο ομάδες και, αφού επιβεβαιώθηκε ότι οι μεταβλητές δεν είναι ανεξάρτητες, επέβαλαν μια διόρθωση στην διμεταβλητή Poisson κατανομή επιτρέποντας την εξάρτηση μεταξύ των γκολ. Επίσης αντιμετώπισαν τη δυναμική φύση των ικανοτήτων των ομάδων χρησιμοποιώντας δεδομένα από το κοντινό παρελθόν για τις επιδόσεις της ομάδας έτσι ώστε πιο πρόσφατα αποτελέσματα να επηρεάζουν την εκτιμώμενη παράμετρο δύναμη περισσότερο, παρά αποτελέσματα πολύ παλαιά στο παρελθόν.

Οι Καρλής και Ντζούφρας (2003) βελτίωσαν το διμεταβλητό Poisson μοντέλο με μια πιο περίπλοκη διατύπωση της συνάρτησης πιθανότητας, αυξάνοντας την πιθανότητα ισοπαλίας, για να μετρήσει το γεγονός ότι οι ισοπαλίες παρατηρούνται πιο συχνά στην πραγματικότητα από ότι στο ανεξάρτητο Poisson μοντέλο. Το νέο μοντέλο προσομοιώνει καλύτερα τις ισοπαλίες, ενώ επιτρέπει υπερδιασπορά και συσχέτιση μεταξύ των μεταβλητών και άρα είναι κατάλληλο για ένα πλήθος εφαρμογών (Baio & Blangiardo, 2010).

Στην συνέχεια αναπτύχθηκε νέο μοντέλο, που βασίζεται στην κατανομή Schellam, το οποίο υπολογίζει τις πιθανότητες για την διαφορά των γκολ. Το 2010 ξεκίνησε η εποχή πιο σύνθετων μοντέλων, με πιο δυνατούς υπολογιστές, και κύριο χαρακτηριστικό ότι οι παράμετροι των μοντέλων αλλάζουν δυναμικά από αγώνα σε αγώνα.

2.4 Κατανομές για τα γκολ

Σύμφωνα με τον Moroney (1956) ο αριθμός των γκολ που βάζουν οι αντίπαλες ομάδες δεν ταιριάζει με Poisson κατανομή αλλά περιγράφεται καλύτερα από μια τροποποίηση της Poisson κατανομής. Οι Reep et al (1971) αργότερα, χρησιμοποιώντας δεδομένα από το αγγλικό ποδόσφαιρο, επιβεβαίωσαν ότι η τροποποημένη Poisson είναι η αρνητική διωνυμική.

Η αρνητική διωνυμική έχει το πλεονέκτημα ότι εφαρμόζεται ανεξάρτητα από την ποιότητα των αντίπαλων ομάδων, δηλαδή η δύναμη μιας ομάδας μπορεί να αλλάζει από παιχνίδι σε παιχνίδι. Ο Maher επεσήμανε ότι η αρνητική διωνυμική μπορεί να προκύψει από την Poisson με διαφορετικό μέσο για κάθε ομάδα. Σύμφωνα με τον Maher (1982), << *Η τυχαιότητα κυριαρχεί σε ένα αγώνα. Χαμένες ευκαιρίες, ύποπτες αβέβαιες αποφάσεις, δοκάρια μπορούν να επιδράσουν στο σκορ δραστικά. Ωστόσο για μεγάλο αριθμό αγώνων η τύχη δεν παίζει τόσο μεγάλο ρόλο. Οι ομάδες δεν είναι ίδιες, κάθε μια έχει την δική της ποιότητα και η καλύτερη έχει μεγαλύτερη πιθανότητα να κερδίσει και να βάλει πολλά γκολ. Χρησιμοποιώντας δεδομένα από όλο το πρωτάθλημα ή ένα μέρος της σεζόν αυτά τα εγγενή χαρακτηριαστικά των ομάδων σε ένα πρωτάθλημα μπορούν να προκύψουν χρησιμοποιώντας για παράδειγμα τη μέγιστη πιθανότητα εκτίμησης ή μεθοδολογία γραμμικών μοντέλων* >>.

Η υπόθεση ότι η κατανομή Poisson ταιριάζει ως κατανομή για γκολ μπορεί να αμφισβητηθεί. Η Poisson κατανομή χρησιμοποιείται για γεγονότα που συμβαίνουν τυχαία και σε ένα σταθερό ρυθμό σε παρατηρούμενη χρονική περίοδο. Σύμφωνα με τους Καρλή και Ντζούφρα (2000) αυτό είναι ισοδύναμο με την υπόθεση ότι η ικανότητα σκοραρίσματος της ομάδας είναι σταθερή σε όλη τη σεζόν. Αυτή η υπόθεση δεν ισχύει στην πραγματικότητα διότι η ικανότητα όπως και η σύνθεση, η φυσική κατάσταση και η τακτική κάθε ομάδας αλλάζει από παιχνίδι σε παιχνίδι. Η υπόθεση της μεταβολής της ικανότητας της ομάδας οδηγεί στην

αρνητική διωνυμική ως την καλύτερη επιλογή και χρησιμοποιείται ευρεώς ως εναλλακτικό μοντέλο της Poisson κατανομής. Η αρνητική διωνυμική μπορεί να παραχθεί από την απλή Poisson κατανομή υποθέτοντας ότι η παράμετρός της μεταβάλλεται σύμφωνα με τη Γάμμα κατανομή.

2.4.1 Poisson κατανομή

Στη στατιστική η κατανομή Poisson είναι μια διακριτή συνάρτηση κατανομής και εκφράζει την πιθανότητα ενός δεδομένου αριθμού γεγονότων που συμβαίνουν σε ένα σταθερό χρονικό διάστημα. Σύμφωνα με τον Hilbe (2011) για να εφαρμοστεί η κατανομή Poisson πρέπει να πληρούνται οι εξής συνθήκες:

α) Οι τιμές της τυχαίας μεταβλητής πρέπει να είναι φυσικοί αριθμοί 0, 1, 2, 3,...
β) Τα γεγονότα να συμβαίνουν τυχαία, δηλαδή ανεξάρτητα της τελευταίας χρονικής στιγμής εμφάνισης του γεγονότος.
γ) Ο μέσος όρος εμφάνισης των γεγονότων λ πρέπει να είναι σταθερός σε όλα τα χρονικά διαστήματα, δηλαδή τα γεγονότα να συμβαίνουν με σταθερό ρυθμό.

Η κατανομή Poisson έχει την παράμετρο λ που δηλώνει τη μέση τιμή αριθμού εμφανίσεων ενός γεγονότος. Αν τα γεγονότα συμβαίνουν σε ένα σταθερό χρονικό διάστημα, με σταθερό μέσο ρυθμό λ και ανεξάρτητα από το χρόνο με το τελευταίο συμβάν, τότε η πιθανότητα να συμβούν κ γεγονότα είναι $f(\kappa, \lambda) = P(X= \kappa) = \frac{\lambda^{\kappa} \cdot e^{-\lambda}}{\kappa!}$. Στην κατανομή Poisson η μέση τιμή ισούται με την διακύμανση.

Πίνακας 1: Κατανομή Poisson

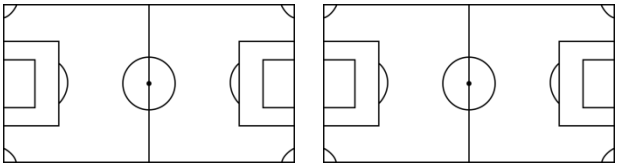
Συνάρτηση πιθανότητας Poisson	Παράμετρος	Μέση τιμή	Διακύμανση
$\frac{\lambda^{\kappa} \cdot e^{-\lambda}}{\kappa!}$	λ	λ	λ

Εφαρμόζοντας σε μια σειρά από αγώνες, το γεγονός είναι το περιστατικό του γκολ, και κάθε παιχνίδι αντιπροσωπεύει μια μονάδα χρονικό διάστημα. Για την Poisson κατανομή η συνθήκη γ απαιτεί ότι το ποσοστό σκοραρίσματος γκολ πρέπει να είναι το ίδιο για όλα τα παιχνίδια.

Γεγονός = Γκολ

Σταθερό χρονικό διάστημα = Παιχνίδι 90 λεπτά

Ρυθμός γεγονότων σταθερός για όλα τα παιχνίδια = λ

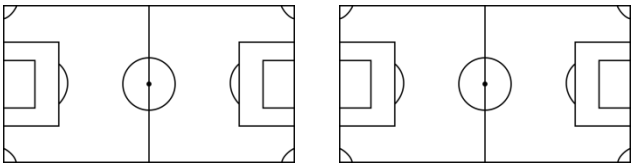


Παιχνίδι 1 = 90’ Παιχνίδι 2 = 90’ .....

Ρυθμός λ Ρυθμός λ

2.4.2 Αρνητική διωνυμική κατανομή

Αν υπάρχει ένα μοντέλο όπου ο μέσος ρυθμός λ επιτρέπεται να μεταβάλλεται από παιχνίδι σε παιχνίδι αλλά μέσα σε κάθε δοθέντα χρονικό διάστημα τα γεγονότα συμβαίνουν τυχαία με σταθερό μέσο ρυθμό, τότε η αναμενόμενη κατανομή των γεγονότων θα είναι αρνητική διωνυμική αντί για Poisson. Ουσιαστικά το λ μπορεί να θεωρηθεί ως τυχαία μεταβλητή και μπορεί να δειχτεί ότι η συνθήκη για μια ακριβή αρνητική διωνυμική είναι το λ να έχει Γάμμα κατανομή (Pollard, 1985). Στο ποδόσφαιρο γίνεται η πιο ρεαλιστική υπόθεση ότι το ποσοστό μέσος ρυθμός σκοραρίσματος μεταβάλλεται από παιχνίδι σε παιχνίδι, αλλά κατά τη διάρκεια κάθε αγώνα τα γκολ παραμένουν να συμβαίνουν τυχαία.



Παιχνίδι 1 = 90’ Παιχνίδι 2 = 90’ .....

Ρυθμός λ_1 Ρυθμός λ_2

Η αρνητική διωνυμική μπορεί να γίνει κατανοητή από την μελέτη της απλής διωνυμικής. Η διωνυμική εκφράζει τον αριθμό των επιτυχιών σε n ανεξάρτητες δοκιμές ενός πειράματος Bernoulli. Για παράδειγμα στην ρίψη ενός ζαριού 20 φορές, η πιθανότητα να έρθουν 4 άσσοι σε

20 δοκιμές είναι $P(X=4) = \binom{20}{4} \cdot \rho^4 \cdot (1-\rho)^{20-4}$ με πιθανότητα επιτυχίας ρ = $\frac{1}{6}$, X= 0, 1, 2,,..., 20. Η αρνητική διωνυμική μοιάζει με την διωνυμική μόνο που οι δοκιμές μπορεί να είναι απεριόριστες. Ενώ στην διωνυμική η τυχαία μεταβλητή μπορεί να πάρει πεπερασμένες τιμές X= 1, 2,..., n, στην αρνητική διωνυμική το X δεν είναι πεπερασμένο αλλά το πείραμα μπορεί να επαναλαμβάνεται απεριόριστα μέχρι να συμβούν r επιτυχίες, X = r, r+1, r+2, r+3, ... όπου r ο αριθμός των επιτυχιών.

Η αρνητική διωνυμική κατανομή είναι μια διακριτή συνάρτηση κατανομής τυχαίας μεταβλητής, και περιγράφει ένα τυχαίο πείραμα με δύο ισοπίθανα αποτελέσματα επιτυχία αποτυχία και πιθανότητα επιτυχίας ρ που επαναλαμβάνεται μέχρι να πραγματοποιηθούν r επιτυχίες. Αυτό που ενδιαφέρει είναι ο αριθμός των δοκιμών x μέχρι να εμφανιστεί η r επιτυχία. Η αντίστοιχη πιθανότητα εώς ότου συμβούν r επιτυχίες και x – r αποτυχίες σε ανεξάρτητα πειράματα με πιθανότητα επιτυχίας ρ είναι $P(X=x; r, p) = \binom{x-1}{r-1} \cdot p^r \cdot (1-p)^{x-r}$**.** Το πείραμα επαναλαμβάνεται x φορές, από τις οποίες η τελευταία είναι η επιτυχία. Η τιμή του x είναι ο συνολικός αριθμός δοκιμών ελπίζοντας να ολοκληρωθεί το πείραμα και r ο αριθμός των επιτυχιών.

Οι συνθήκες της αρνητικής διωνυμικής είναι οι ακόλουθες:
α) Είναι ένα πείραμα Bernoulli που σημαίνει ότι υπάρχουν δύο αποτελέσματα αποτυχία-επιτυχία.
β) Η πιθανότητα επιτυχίας είναι σταθερή και ίση με ρ.
γ) Το πείραμα επαναλαμβάνεται για ανεξάρτητες δοκιμές που σημαίνει ότι το αποτέλεσμα μιας δοκιμής δεν έχει καμιά επίδραση στο αποτέλεσμα της επόμενης δοκιμής.
δ) Οι τιμές της τυχαίας μεταβλητής είναι διακριτές.
ε) Σε αντίθεση με την απλή διωνυμική όπου η τυχαία μεταβλητή μπορεί να πάρει τιμές X=1, 2,..., n, στην αρνητική το X= r, r+1, r+2 ,r+3, ... μπορεί να πάρει απεριόριστα μεγάλες τιμές μέχρι να επιτευχθούν r επιτυχίες.

Σύμφωνα με τον Pollard (1985) αν r είναι ο αριθμός των γκολ που βάζει μια ομάδα σε ένα αγώνα με πιθανότητα ρ ανά γκολ πριν κ χαμένα γκολ με πιθανότητα q συμβούν, τότε η αρνητική διωνυμική δίνει $P(X = r + k) = \binom{r+k-1}{r-1} \cdot p^r \cdot q^k$, r = 0, 1, 2,… όπου r > 0 και 0 < ρ < 1 είναι οι παράμετροι της κατανομής οι οποίες και οι δύο μπορούν να εκφραστούν σε όρους του μέσου και διακύμανσης του X **.** Η αρνητική διωνυμική έχει $\mu = \frac{\mathbf{r \cdot (1-p)}}{\mathbf{p}}$ και $\sigma^2 = \frac{\mathbf{r \cdot (1-p)}}{\mathbf{p^2}}$.

Προσαρμόζοντας τα δεδομένα απαιτείται $\frac{\mu}{\sigma^2} = p \leq 1$ και $\frac{\mu \cdot p}{1-p} = r$, όπου υπολογίζεται το r στον πλησιέστερο ακέραιο. Άρα στην αρνητική διωνυμική πρέπει $\mu \leq \sigma^2$.

2.4.3 Υπόθεση Poisson

Σύμφωνα με τους Καρλή και Ντζούφρα (2008) η υπόθεση ότι η Poisson κατανομή ταιριάζει για να περιγράψει τα γκολ αμφισβητείται, διότι σε πολλά πρωταθλήματα παρατηρείται ότι η δειγματική διακύμανση είναι μεγαλύτερη του μέσου όρου. Η Poisson κατανομή αντίθετα προϋποθέτει ότι η διακύμανση είναι ίση με το μέσο, δηλαδή για να ταιριάξουν τα δεδομένα καλά χρησιμοποιώντας Poisson κατανομή θα πρέπει $\mu \cong \sigma^2$. Σύμφωνα με τον Hilbe (2011) η συνθήκη $\mu \cong \sigma^2$ ισχύει μόνο όταν τα γεγονότα συμβαίνουν μέσα σε μια περίοδο παρατηρήσεων με ένα σταθερό ρυθμό, που σημαίνει ότι ένα γεγονός είναι ισοπίθανο σε κάθε σημείο της περιόδου. Αυτό δεν ισχύει για τα γκολ διότι η υπόθεση της σταθερής πιθανότητας ανά μονάδα χρόνου σκοραρίσματος γκολ δεν είναι έγκυρη. Σε αυτή την περίπτωση θα υπάρχει ανομοιογένεια στα δεδομένα και το μοντέλο Poisson θα έχει υπερδιασπορά, που υποδεικνύεται αν η διακύμανση είναι μεγαλύτερη από τον μέσο $\sigma^2 > \mu$. Το γεγονός ότι πολλές φορές παραβιάζεται η υπόθεση, η διακύμανση είναι ίση με τον μέσο, είναι ο λόγος που η αρνητική διωνυμική έχει πιο αποτελεσματική και ευρεία χρήση από την Poisson, σε μοντέλα καταμέτρησης δεδομένων. Ωστόσο αποδεικνύεται ότι οι κατανομές δεν παρουσιάζουν μεγάλες διαφορές, ενώ εξαιτίας της περίπλοκης φύσης της αρνητικής διωνυμικής πολλοί ερευνητές χρησιμοποιούν με ασφάλεια το Poisson μοντέλο.

2.5 Ανεξαρτησία

2.5.1 Εξάρτηση και συσχέτιση

Στην στατιστική η εξάρτηση είναι οποιαδήποτε στατιστική σχέση μεταξύ δύο τυχαίων μεταβλητών ή δύο συνόλων δεδομένων. Γνωστά παραδείγματα εξαρτημένων φαινόμενων περιλαμβάνουν τη συσχέτιση μεταξύ των φυσικών φαινοτύπων των γονέων και των απογόνων τους, καθώς και τη συσχέτιση μεταξύ της ζήτησης για ένα προϊόν και την τιμή του. Οι τυχαίες μεταβλητές X και Y καλούνται ανεξάρτητες αν για οποιαδήποτε υποσύνολα A και B του συνόλου των πραγματικών αριθμών, ισχύει ότι $P\{X \in A, Y \in B\} = P\{X \in A\} \cdot P\{Y \in B\}$. Όταν οι X και Y είναι διακριτές τυχαίες μεταβλητές, η συνθήκη ανεξαρτησίας είναι ισοδύναμη με τη σχέση $P(x,y) = p_x(x) \cdot p_y(y)$, ή με τη σχέση $P[X = x, Y = y] = P[X = x] \cdot P[Y = y]$, για κάθε x, y. Διαισθητικά, οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες όταν η γνώση της κατανομής της μίας δεν επηρεάζει την κατανομή της άλλης. Αν δύο τυχαίες μεταβλητές δεν είναι ανεξάρτητες, τότε καλούνται εξαρτημένες (Δημητράκος, 2010).

Υπάρχει διαφορά μεταξύ των εννοιών συσχέτιση και εξάρτηση στην στατιστική. Ο όρος συσχέτιση χρησιμοποιείται για ένα συγκεκριμένο τύπο συσχέτισης, την συσχέτιση Pearson, η οποία αναφέρεται στο βαθμό που δύο μεταβλητές σχετίζονται γραμμικά. Ο δειγματικός συντελεστής συσχέτισης γράφεται $r_{xy}= \frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{\sqrt{\sum_{i=1}^{n}(xi-\bar{x})^2(yi-\bar{y})^2}}$ όπου x και y είναι ο δειγματικός μέσος των X και Y . Η εξάρτηση αντίθετα είναι οποιαδήποτε στατιστική σχέση μεταξύ δύο τυχαίων μεταβλητών ή δύο συνόλων δεδομένων και αναφέρεται σε οποιαδήποτε κατάσταση στην οποία τυχαίες μεταβλητές δεν πληρούν μια μαθηματική κατάσταση πιθανολογικής ανεξαρτησίας. Ωστόσο η συσχέτιση Pearson συνδέεται με την εξάρτηση. Αποδεικνύεται ότι αν δύο μεταβλητές είναι συσχετισμένες, τότε είναι εξαρτημένες. Όμως το αντίστροφο δεν ισχύει δηλαδή μπορεί να είναι εξαρτημένες και να μην είναι συσχετισμένες. Αυτό συμβαίνει διότι ο συντελεστής συσχέτισης ανιχνεύει μόνο γραμμική σχέση μεταξύ των μεταβλητών. Τέλος στην ειδική περίπτωση, όταν X και Y είναι από κοινού κανονικές, το ότι δε συσχετίζονται είναι ισοδύναμο με την ανεξαρτησία.

2.5.2 Σημασία ανεξαρτησίας

Η υπόθεση ανεξαρτησίας στο σκορ σημαίνει ότι τα γκολ των ομάδων δεν σχετίζονται μεταξύ τους, δηλαδή τα γκολ που βάζει η μια ομάδα δεν πρέπει να επηρεάζουν με κάποιο τρόπο τα γκολ που βάζει η αντίπαλη ομάδα. Αυτή η υπόθεση είναι δύσκολο να υποστηριχθεί διότι η εξάρτηση στα δεδομένα είναι περίπλοκη.

Μερικές φορές τα δεδομένα είναι συσχετισμένα, όπως όταν υπάρχει το ίδιο άτομο πολλές φορές ή μέλη της ίδιας οικογένειας ή μαθητές που διδάσκονται από τους ίδιους καθηγητές. Η έννοια της εξάρτησης γίνεται κατανοητή με το παρακάτω παράδειγμα. Έστω ότι ρωτήθηκαν φίλαθλοι στην Θεσσαλονίκη για το ποια ομάδα ποδοσφαίρου υποστηρίζουν οι Έλληνες φίλαθλοι. Έστω ότι ρωτήθηκαν 10 άτομα και οι 7 ήταν ΠΑΟΚ. Τότε η πιθανότητα να είναι ΠΑΟΚ ένας πολίτης στην Θεσσαλονίκη είναι 70%. Στην πραγματικότητα η πιθανότητα να είναι ΠΑΟΚ κάποιος φίλαθλος στην Ελλάδα είναι πολύ μικρότερη από 70%. Αν επαναληφθεί το ίδιο πείραμα στην Θεσσαλονίκη διπλασιάζοντας τους ερωτώμενους, και ενώ κάποιος θα πίστευε ότι αυξάνοντας τα δεδομένα το αποτέλεσμα θα ήταν πιο έγκυρο, η πιθανότητα να είναι κάποιος ΠΑΟΚ θα είναι πάλι 70%. Βρέθηκε το ίδιο αποτέλεσμα διότι έγινε αντιγραφή των δεδομένων. Αν όμως η έρευνα γινόταν από ένα άλλο τυχαίο δείγμα, και όχι στη Θεσσαλονίκη, όπου η συντριπτική πλειοψηφία των φιλάθλων είναι ΠΑΟΚ, τα αποτελέσματα θα ήταν διαφορετικά. Άρα ο διπλασιασμός των δεδομένων δεν οδηγεί πάντα σε αξιόπιστα αποτελέσματα. Η πιθανότητα να υποστηρίζεται μια ομάδα από το 70% των φιλάθλων σε ανεξάρτητα δεδομένα

είναι τελείως διαφορετικό από το να βρεθεί 70% με αντιγραφή των δεδομένων. Και αντιγραφή των δεδομένων μπορεί να γίνει όταν τα δεδομένα είναι εξαρτημένα. Σύμφωνα με ερευνητές, από όλες τις υποθέσεις που γίνονται στην στατιστική, όπως είναι η ετεροδιασκεδαστικότητα, η κανονικότητα, η ανεξαρτησία, η γραμμικότητα, η πιο σημαντική είναι αυτή της ανεξαρτησίας γιατί υπάρχει κίνδυνος να αλλάξουν τεχνητά τα δεδομένα και να προκύψουν λανθασμένες πιθανότητες.

2.5.3 Υπόθεση ανεξαρτησίας στα γκολ

Ο Maher (1982) θεώρησε ότι τα γκολ που βάζουν οι δύο ομάδες, στο ανεξάρτητο Poisson μοντέλο, είναι μεταβλητές ανεξάρτητες. Παρ'όλα αυτά αποδείχτηκε ότι, αν και μικρή, υπάρχει μια συσχέτιση μεταξύ των δύο ποσοτήτων. Ο Maher διόρθωσε το αρχικό του μοντέλο σε διμεταβλητό Poisson για να αντιμετωπίσει το πρόβλημα. Στη συνέχεια άλλοι ερευνητές όπως οι Dixon and Coles (1997) έδειξαν ότι η εξάρτηση είναι πιο σύνθετη στα δεδομένα που εξέτασαν. Βρήκαν μεγάλο αριθμό αποτελεσμάτων 0-0 και 1-1 από ότι θα αναμενόταν και λιγότερα αποτελέσματα 0-1, 1-0. Ανέπτυξαν ένα μοντέλο προτείνοντας παραμέτρους που αποτελεσματικά αυξάνουν την πιθανότητα χαμηλών σκορ και μειώνουν τις πιθανότητες αποτελεσμάτων 1-0 και 0-1. Οι Καρλής και Ντζούφρας (2003) επιχείρησαν να ενσωματώσουν τη συσχέτιση σε πιο βελτιωμένα μοντέλα που μετράνε τον παράγοντα εξάρτησης. Χρησιμοποίησαν διμεταβλητό Poisson μοντέλο όπου τα γκολ X και Y μοντελοποιούνται με τρεις Poisson τυχαίες μεταβλητές Z_1, Z_2, Z_3 με μέσους λ_1, λ_2, λ_3, αντίστοιχα όπου $X= Z_1+ Z_3$ και $Y= Z_2 + Z_3$, ενώ η κοινή συνάρτηση πιθανότητας δίνεται από τον τύπο $P(X=x, Y=y) = \exp(-\lambda_1-\lambda_2-\lambda_3)\cdot \sum_{\kappa=0}^{\min(x,y)} \frac{\lambda_1\cdot\lambda_2\cdot\lambda_3}{(x-k)!\cdot(y-k)!}$.

Οι McHale and Scarf (2007) πρότειναν μοντέλα που βασίζονται σε παραστάσεις copula. Η χρήση της copula συνδέει τις οριακές κατανομές των γκολ με τη διακριτή διμεταβλητή κατανομή. Η copula παράσταση επιτρέπει εξάρτηση στη διμεταβλητή κατανομή για να μοντελοποιηθεί με ένα ευέλικτο τρόπο καθορίζοντας μια κατάλληλη οικογένεια συναρτήσεων copula και προσαρμόζοντας αυτό στα διμεταβλητά δεδομένα χρησιμοποιώντας τη μέθοδο της μέγιστης πιθανότητας. Σύμφωνα με τους Boshnakov et al (2017) μια παράσταση copula είναι μια πολυμεταβλητή κατανομή για την οποία όλες οι μονομεταβλητές οριακές κατανομές κατανέμονται ομοιόμορφα στο διάστημα της μονάδας [0,1]. Το πλεονέκτημα της copula προσέγγισης για μοντελοποίηση προέρχεται από το θεώρημα του Sclar (1973) ότι η κοινή αθροιστική κατανομή συνάρτηση F για κάθε ζεύγος τυχαίων μεταβλητών (Y1,Y2) μπορεί να γραφτεί με τον τύπο $F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$, $(y_1, y_2) \in R^*$ όπου F_1, F_2 είναι οι οριακές αθροιστικές κατανομές της συνάρτησης και C είναι μια copula.

Κεφάλαιο 3 Έλεγχος υποθέσεων

Εκτός από την εξαγωγή μιας υποκείμενης κατανομής πιθανοτήτων, ουσιώδες για την ανάλυση δεδομένων είναι τα στατιστικά τεστ και οι υποθέσεις. Τα στατιστικά τεστ είναι ένα εργαλείο για την σωστή ερμηνεία των αποτελεσμάτων της έρευνας και την αποκάλυψη της αλήθειας, ενώ έχουν μεγάλη σημασία διότι μπορούν να αλλάξουν τα συμπεράσματα σε μια μελέτη. Η δύναμη της αληθοφάνειας και της ποιότητας μιας έρευνας εξαρτάται από τη διαδικασία που ακολουθείται, αλλά και την αποδοτικότητα των στατιστικών τεστ. Οι παράγοντες που επηρεάζουν ένα στατιστικό τεστ είναι ο ερευνητικός σκοπός, τα δεδομένα και οι υποθέσεις που χρησιμοποιούνται.

3.1 Στατιστικός έλεγχος υποθέσεων

Ο στατιστικός έλεγχος υποθέσεων είναι μια τεχνική για να αποφασιστεί αν το δείγμα υποστηρίζει επαρκώς μια συγκεκριμένη υπόθεση ή εικασία. Η γενική ιδέα της μεθόδου είναι ότι ορίζεται ως μηδενική υπόθεση H_0 αυτή η οποία αμφισβητείται. Στην συνέχεια εξετάζεται αν τα δεδομένα δίνουν αποδείξεις υπέρ αυτής της υπόθεσης, ή αν αντίθετα συνηγορούν στην απόρριψη της H_0. Αν απορριφθεί η H_0 τότε γίνεται δεκτή η εναλλακτική πρόταση της H_0 που συμβολίζεται H_1. Η μη απόρριψη της H_0 μπορεί να είναι ένα καλό αποτέλεσμα για να θεωρηθεί αληθής η μηδενική υπόθεση, ή μπορεί να σημαίνει ότι δεν υπάρχουν επαρκή δεδομένα για να αποδειχθεί κάτι απορρίπτοντας την μηδενική υπόθεση (Παπαδόπουλος 2015).

Ο έλεγχος υποθέσεων διαδόθηκε στις αρχές του 20ου αιώνα αλλά οι πρώιμες μορφές του υπήρχαν ήδη από παλαιότερες αναφορές. Ο Laplace το 1778 σύγκρινε τα ποσοστά γεννήσεων αγοριών και κοριτσιών σε πολλές Ευρωπαικές πόλεις και δήλωσε ότι << *Είναι φυσικό να συμπεραίνουμε ότι αυτές οι πιθανότητες είναι σχεδόν στην ίδια αναλογία* >>. Η μηδενική υπόθεση του Laplace ήταν ότι τα ποσοστά των γεννήσεων αγοριών και κοριτσιών πρέπει να είναι ίσα δεδομένης της διαίσθησής μας.

3.1.1 Επίπεδο σημαντικότητας α

Η μηδενική υπόθεση, σε αντίθεση με την εναλλακτική της, αφορά σε μια υπόθεση που εκ πρώτης όψεως φαίνεται ότι ισχύει, αν και δοκιμάζεται. Μπορεί η εναλλακτική να είναι τελικά η αληθής, όμως η μέθοδος είναι κατασκευασμένη έτσι ώστε ο κίνδυνος απόρριψης της μηδενικής υπόθεσης, όταν στην πραγματικότητα είναι αληθής, να είναι μικρός. Για αυτό η μηδενική υπόθεση είναι αυτή που από τη θεωρία ή την αρχική αντίληψη φαίνεται να είναι σωστή. Ο κίνδυνος η μηδενική υπόθεση να απορριφθεί, ενώ στην πραγματικότητα είναι αληθής, αναφέρεται ως επίπεδο σημαντικότητας του τεστ και συμβολίζεται με α. Μικρή τιμή του α σε ένα τεστ σημαίνει ότι μάλλον η απόδειξη είναι σωστή όταν απορρίπτεται η μηδενική υπόθεση.

Επίσης υπάρχει ο κίνδυνος να γίνει δεκτή η μηδενική υπόθεση όταν είναι εσφαλμένη και αυτός ο κίνδυνος δεν επιλέγεται από τον χρήστη αλλά καθορίζεται από το μέγεθος της πραγματικής απόκλισης και λέγεται σφάλμα δεύτερου βαθμού (Παπαδόπουλος, 2015).

Ένας στατιστικός έλεγχος απαιτεί ένα ζεύγος υποθέσεων

H_0: Μηδενική υπόθεση

H_1: Εναλλακτική υπόθεση

3.1.2 Παρατηρούμενο επίπεδο σημαντικότητας p-value

Η τιμή του ρ-value είναι ένα αντικειμενικό μέτρο για την στατιστική σημασία του αποτελέσματος σε μια δοκιμή στατιστικής υπόθεσης. Ένα αποτέλεσμα λέγεται ότι είναι στατιστικά σημαντικό εάν επιτρέπει να απορριφθεί η μηδενική υπόθεση. Το p-value δείχνει πόσο πιθανό είναι το αποτέλεσμα, θεωρώντας ότι ισχύει η μηδενική υπόθεση, ενώ όσο μικρότερο είναι τόσο λιγότερο πιθανό είναι να ισχύει η μηδενική υπόθεση. Η τιμή ρ-value ονομάζεται επίσης παρατηρούμενο επίπεδο σημαντικότητας και η σύγκρισή του με το επίπεδο σημαντικότητας α δείχνει αν απορρίπτεται η H_0. Συγκεκριμένα αν η τιμή p-value είναι μικρότερη από το α, απορρίπτεται η μηδενική υπόθεση. Διαφορετικά, αν p-value $\geq$ α , δεν απορρίπτεται η μηδενική υπόθεση H_0. Η στατιστική σημαντικότητα είναι θεμελιώδης για την δοκιμή στατιστικής υπόθεσης. Η απόρριψη της H_0 και το γεγονός ότι το αποτέλεσμα είναι σταστιστικά σημαντικό σημαίνει επίσης για τον ερευνητή ότι το αποτέλεσμα που προέκυψε δεν είναι προϊόν σφάλματος στη δειγματοληψία, αλλά το δείγμα αντανακλά στην πραγματικότητα τα χαρακτηριστικά του πληθυσμού (Ziliak & McCloske, 2008).

3.2 Στατιστική δοκιμή

Στατιστικό στοιχείο είναι κάθε ποσότητα που υπολογίζεται από τιμές σε ένα δείγμα και θεωρείται για στατιστικό σκοπό. Ο σκοπός μπορεί να αφορά την εκτίμηση μιας παραμέτρου, την περιγραφή ενός δείγματος ή την αξιολόγηση μιας υπόθεσης, ενώ για κάθε μια περίπτωση χρησιμοποιείται και άλλο στατιστικό (Sybil, 1994).

Πίνακας 2: Σκοπός στατιστικού και χρήση

Σκοπός στατιστικού	Χρήση
Εκτίμηση μιας παραμέτρου (πχ εκτίμηση μέσου όρου πληθυσμού)	Χαρακτηριστικό πληθυσμού -εκτιμητής (π.χ. μέσος όρος δείγματος)

Περιγραφή δείγματος	Περιγραφική στατιστική
Έλεγχος υποθέσεων	Στατιστική δοκιμή

Υπάρχει μια ποικιλία συναρτήσεων που χρησιμοποιούνται για τον υπολογισμό των στατιστικών. Ορισμένα στατιστικά είναι τα εξής: Μέσος δείγματος, διάμεσος, διακύμανση, τυπική απόκλιση, τεταρτημόρια, στατιστικά δοκιμών όπως t στατιστικό, χ τετράγωνο στατιστικό, f στατιστικό, δείγμα ροπών.

Η στατιστική δοκιμή είναι ένας αριθμός σύνοψη ενός συνόλου δεδομένων που μειώνει τα δεδομένα σε μια τιμή και μπορεί να χρησιμοποιηθεί στον έλεγχο υποθέσεων. Η τιμή αυτή ποσοτικοποιεί εντός των παρατηρούμενων δεδομένων συμπεριφορές που θα ξεχωρίσουν την μηδενική από την εναλλακτική υπόθεση. Σημαντική ιδιότητα της στατιστική δοκιμής είναι ότι η κατανομή δειγματοληψίας στην μηδενική υπόθεση πρέπει να είναι υπολογίσιμη, έστω κατά προσέγγιση, ώστε να μπορεί να υπολογιστεί το p–value. Πολλά στατιστικά στοιχεία μπορούν να χρησιμοποιηθούν τόσο ως στοιχεία δοκιμής όσο και ως περιγραφικά στοιχεία, δηλαδή και σε έλεγχο υποθέσεων και σε περιγραφική στατιστική. Ωστόσο σε έλεγχο υποθέσεων είναι πιο δύσκολη η ερμηνεία τους.

3.2.1 Στατισικά τεστ- δοκιμές

Ο παρακάτω πίνακας παρουσιάζει τις διαφορές μεταξύ των πιο κοινών δοκιμών και τις συνθήκες κάτω από τις οποίες πρέπει να χρησιμοποιούνται.

Πίνακας 3: Διαφορές μεταξύ των στατιστικών δοκιμών

Στατιστικά τεστ	Εφαρμογές
Δοκιμές ενός δείγματος	Ένα δείγμα συγκρίνεται με τον πληθυσμό από μια υπόθεση
Δοκιμές δύο δειγμάτων	Δύο δείγματα συγκρίνονται, όπου το ένα είναι συνήθως δείγμα ελέγχου με ένα ελεγχόμενο πείραμα
Ζευγαρωμένα τεστ	Σύγκριση δύο δειγμάτων, όπου η

	διαφορά των μελών γίνεται δείγμα. Ο μέσος όρος των διαφορών συγκρίνεται με το μηδέν
Δοκιμές Z	Σύγκριση μέσων υπό αυστηρές συνθήκες όσον αφορά την κανονικότητα και μια γνωστή τυπική απόκλιση. Χρησιμοποιείται για να επικυρώσει μια υπόθεση που δηλώνει ότι το δείγμα ανήκει στον ίδιο πληθυσμό
t-τεστ	Σύγκριση μέσων δύο δειγμάτων, προϋποθέτει κανονική κατανομή του δείγματος. Όταν δεν είναι γνωστές οι παράμετροι πληθυσμού (μέσος όρος και τυπική απόκλιση)
Τεστ F (ανάλυση διακύμανης ANOVA)	Όταν οι ομαδοποιήσεις δεδομένων ανά κατηγορία έχουν νόημα. Για σύγκριση τριών ή περισσότερων δειγμάτων με ένα μόνο τεστ
Τεστ χ-τετράγωνο	Σύγκριση κατηγορικών μεταβλητών
Δοκιμή χ-τετράγωνο για διακύμανση	Για να προσδιοριστεί εάν ένας κανονικός πληθυσμός έχει μια καθορισμένη διακύμανση. Η μηδενική υπόθεση είναι ότι ισχύει
Δοκιμή χ-τετράγωνο για ανεξαρτησία	Για να αποφασιστεί εάν δύο μεταβλητές συνδέονται ή είναι ανεξάρτητες.Οι μεταβλητές είναι κατηγορικές και όχι αριθμητικές. Οι αριθμοί που χρησιμοποιούνται στον υπολογισμό είναι οι παρατηρούμενες και αναμενόμενες συχνότητες εμφάνισης από τον πίνακα έκτακτης ανάγκης. Η μηδενική υπόθεση

	είναι ότι οι μεταβλητές είναι ανεξάρτητες
Δοκιμή χ-τετράγωνο καλής προσαρμογής	Για τον προσδιορισμό της επάρκειας των καμπυλών που ταιριάζουν στα δεδομένα. Η μηδενική υπόθεση είναι ότι η προσαρμογή της καμπύλης είναι επαρκής. Είναι σύνηθες να προσδιορίζονται σχήματα καμπυλών για την ελαχιστοποίηση του μέσου τετραγώνου σφάλματος, επομένως είναι σκόπιμο ο υπολογισμός της καλής προσαρμογής να αθροίζει τα τετράγωνα σφάλματα

Ο τρόπος χρήσης της μηδενικής υπόθεσης αλλάζει ανάλογα με την δοκιμή, για παράδειγμα:

Z-τεστ

H_0: Ο μέσος όρος του δείγματος είναι ίδιος με τον μέσο όρο του πληθυσμού.

H_1: Ο μέσος όρος του δείγματος δεν είναι ίδιος με τον μέσο όρο του πληθυσμού.

Ανάλυση διασποράς ANOVA

H_0: Όλα τα ζεύγη δειγμάτων είναι ίδια, δηλαδή όλα τα μέσα δείγματα είναι ίσα.

H_1: Τουλάχιστον ένα ζεύγος δειγμάτων είναι σημαντικά διαφορετικό.

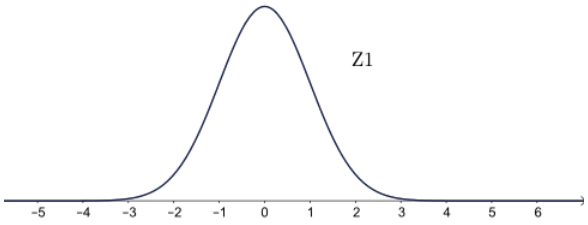
Τεστ χ – τετράγωνο ανεξαρτησίας

H_0: Η μεταβλητή A και η μεταβλητή B είναι ανεξάρτητες.

H_1: Η μεταβλητή A και η μεταβλητή B δεν είναι ανεξάρτητες.

3.2.2 Η κατανομή χ^2

Η κατανομή χ^2 ανακαλύφθηκε, σχεδόν ταυτόχρονα και ανεξάρτητα, από τον Γερμανό μαθηματικό Frederick Robert Helmer το 1875 και αργότερα από τον Karl Pearson το 1900. Η χ^2 κατανομή έχει εξίσωση $X_\kappa^2 = \sum_{i=1}^{k} Z_i^2$, όπου η μεταβλητή Z_i ακολουθεί μια τυπικά κανονική κατανομή $Z_i \sim N(0,1)$. Η τυπική κανονική κατανομή είναι μια καμπύλη, όπως το παρακάνω σχήμα, όπου τα περισσότερα δεδομένα βρίσκονται γύρω από το μηδέν 0, ενώ υπάρχουν λιγότερα δεδομένα Z καθώς οι τιμές μεγαλώνουν ή μικραίνουν και προς τις δύο κατευθύνσεις.

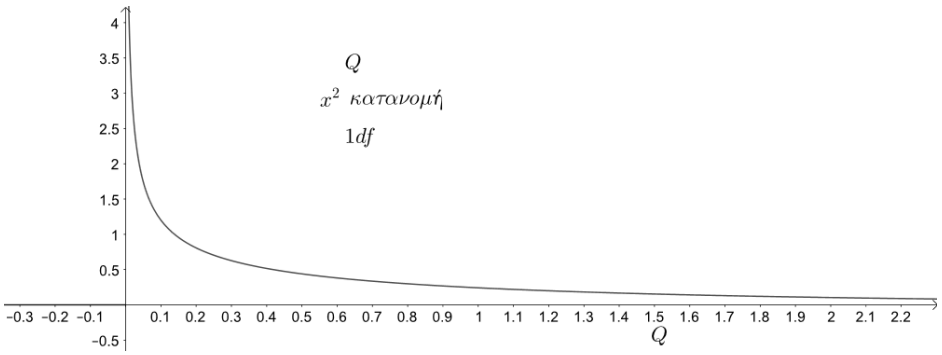


Γράφημα 1: Κανονική κατανομή N(0,1)

Έστω ότι η τυχαία μεταβλητή Z_1 κατανέμεται κανονικά με μέσο 0 και διακύμανση 1, Z_1~ N(0,1). Αν αναζητηθεί η κατανομή των τετραγώνων του Z_1, δηλαδή $Q = Z_1^2$ τότε επειδή οι περισσότερες τιμές του Z_1 κινούνται μεταξύ -1 και 1, οι περισσότερες τιμές του $Q = Z_1^2$ θα είναι πολύ μικρές. Άρα η κατανομή $Q = Z_1^2 \sim X^2$ θα έχει την παρακάτω μορφή.

Πίνακας 4: Κατανομή $Q = Z_1^2 \sim X^2$

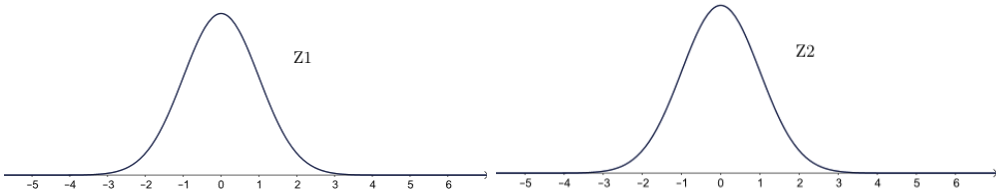
Z_1	- 0,9	- 0,8	0,4	0,8	0,9
$Q = Z_1^2$	0,81	0,64	0,16	0,64	0,81



Γράφημα 2: Κατανομή χ^2 με 1 βαθμό ελευθερίας

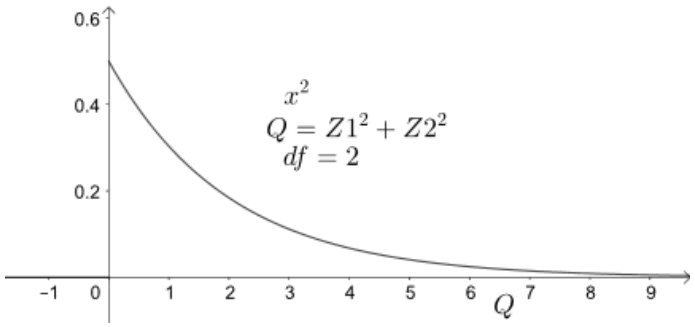
Άρα αν Z_1~N(0,1) ακολουθεί τυπική κανονική κατανομή τότε $Z_1^2 \sim X^2$ θα ακολουθεί κατανομή χ^2. Έχει 1 βαθμό ελευθερίας df =1 και γράφεται $Q = Z_1^2 \sim X^2(1)$ διότι είναι μια η μεταβλητή Z_1 από την κανονική κατανομή που υψώθηκε στο τετράγωνο και το ζητούμενο είναι η πιθανότητα κάθε μιας από αυτές τις δυνητικές τετραγωνικές τιμές Z_1.

Έστω ότι χρησιμοποιούνται δύο διαφορετικές τυχαίες μεταβλητές από την τυποποιημένη κανονική κατανομή, $Z_1 \sim N(0,1)$ και $Z_2 \sim N(0,1)$.



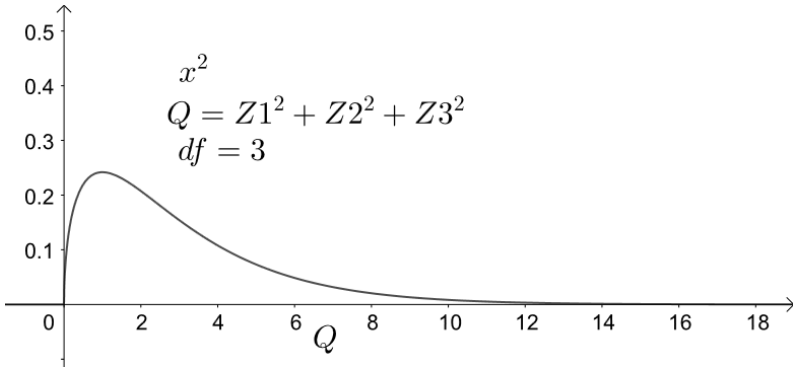
Γράφημα 3: Δύο τυχαίες μεταβλητές από την κανονική κατανομή N(0,1)

Αν αθροιστούν τα τετράγωνα των Z_1 και Z_2, τότε το άθροισμα κατανέμεται ως χ^2 κατανομή $Q = Z_1^2 + Z_2^2 \sim X^2(2)$ με 2 βαθμού ελευθερίας. Η καμπύλη σε αυτή την περίπτωση θα έχει μορφή



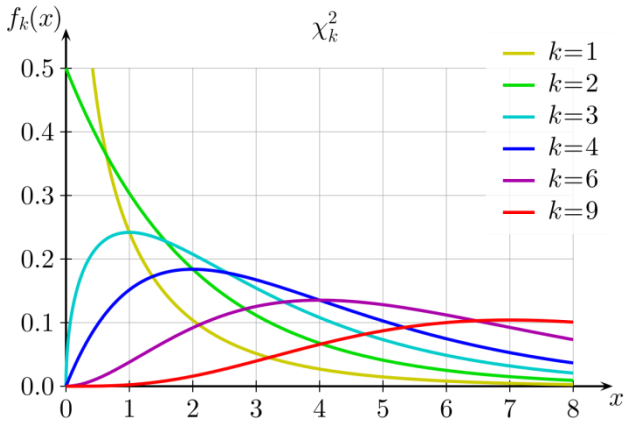
Γράφημα 4: Κατανομή χ^2 με 2 βαθμό ελευθερίας

Αν αθροιστούν τα τετράγωνα τριών μεταβλητών Z_1, Z_2, Z_3 που έχουν τυπική κανονική κατανομή τότε το άθροισμα θα έχει κατανομή χ^2 με 3 βαθμούς ελευθερίας $Q = Z_1^2 + Z_2^2 + Z_3^2 \sim X^2(3)$ και την παρακάτω καμπύλη.



Γράφημα 5: Κατανομή χ^2 με 3 βαθμό ελευθερίας

Άρα η κατανομή χ^2 περιλαμβάνει μια οικογένεια κατανομών χ^2 για διαφορετικούς βαθμούς ελευθερίας κ, όπου κ οι όροι που προσθέτονται. Ο γενικός τύπος είναι $Q = \sum_{i=1}^{k} Z_i^2 \sim X^2(\kappa)$, όπου Z_i κανονικές τυπικές τιμές. Επίσης $E(X_\kappa^2) = \kappa$ και $Var(X_\kappa^2) = 2\kappa$.



Γράφημα 6: Οικογένεια κατανομών χ^2

3.2.3 Η σημασία της χ^2 κατανομής

Η κατανομή χ^2 χρησιμοποιείται σε πολυάριθμες εφαρμογές ή δοκιμές ως η κατάλληλη κατανομή, διότι πολλές φορές τα στατιστικά στοιχεία κατανέμονται κατά προσέγγιση ως χ^2. Για παράδειγμα στην γραμμική παλινδρόμηση το μέσο τετραγωνικό σφάλμα (MSE) $= \frac{\sum(Y_i - \widehat{Y}_i)^2}{n-k-1}$ ακολουθεί χ^2 κατανομή με την παραδοχή ότι τα σφάλματα $Y_i - \widehat{Y}_i$ είναι κανονικά κατανεμημένα. Επίσης η δοκιμή ότι η διακύμανση ενός κανονικά κατανεμημένου πληθυσμού έχει μια δεδομένη τιμή με βάση μια διακύμανση δείγματος γίνεται με δοκιμή χ^2 για να προσδιοριστεί αν η διακύμανση του πληθυσμού ισούται με μια δεδομένη σταθερά, με τη διακύμανση του δείγματος.

Κάθε δοκιμή στην οποία η δειγματοληπτική κατανομή από το στατιστικό αποτέλεσμα είναι χ^2 κατανομή και εφαρμόζεται σε σύνολα κατηγορικών δεδομένων ονομάζεται χ^2 τεστ. Γενικά η δοκιμή αυτή αξιολογεί πόσο πιθανό είναι να προέκυψε τυχαία οποιαδήποτε παρατηρούμενη διαφορά μεταξύ των συνόλων, μέσω της μηδενικής υπόθεσης ότι η συχνότητα ορισμένων γεγονότων ταιριάζει με μια συγκεκριμένη θεωρητική κατανομή.

Οι πιο δημοφιλείς τύποι σύγκρισης που χρησιμοποιείται το χ^2 τεστ είναι το τεστ καλής προσαρμογής και το τεστ ανεξαρτησίας. Το τεστ καλής προσαρμογής καθορίζει αν το δείγμα προσαρμόζεται καλά στον πληθυσμό, ενώ το τεστ ανεξαρτησίας αξιολογεί αν οι παρατηρήσεις

δύο κατηγορικών μεταβλητών που εκφράζονται σε ένα πίνακα έκτακτης ανάγκης είναι ανεξάρτητες μεταξύ τους. Μικρή τιμή του X^2 στατιστικού σημαίνει ότι τα δεδομένα ταιριάζουν, ενώ για μεγάλη τιμή δεν ταιριάζουν. Οι διαφορές των δύο τύπων τεστ μπορούν να εξηγηθούν από την μηδενική υπόθεση και τον τύπο του δείγματος, διότι η ερμηνεία αλλάζει ανάλογα με αυτό.

Πίνακας 5: Διαφορές μεταξύ τεστ καλής προσαρμογής και ανεξαρτησίας

χ^2 χαρακτηριστικό	Τύπος δειγματοληψίας	Ερμηνεία	Μηδενική υπόθεση
Τεστ ανεξαρτησίας	Μονό εξαρτημένο δείγμα	Συσχέτιση μεταξύ μεταβλητών	Καμία συσχέτιση μεταξύ μεταβλητών
Τεστ καταλληλότητας	Δείγμα από πληθυσμό	Διαφορά από πληθυσμό	Καμία διαφορά στην κατανομή μεταξύ δείγματος και πληθυσμού

Και στα δύο τεστ η διαδικασία έχει ως εξής:

1. Υπολογίζεται το στατιστικό X^2 το οποίο μοιάζει με ένα κανονικοποιημένο άθροισμα τετραγωνικών αποκλίσεων μεταξύ παρατηρούμενων και θεωρητικών συχνοτήτων.
2. Προσδιορίζονται οι βαθμοί ελευθερίας df.
3. Επιλέγεται ένα επιθυμητό επίπεδο εμπιστοσύνης (επίπεδο σημαντικότητας, τιμή p ή το αντίστοιχο επίπεδο α) για το αποτέλεσμα της δοκιμής.
4. Συγκρίνεται η τιμή X^2 με την κρίσιμη τιμή από την κατανομή χ^2 με df βαθμούς ελευθερίας, η οποία σε πολλές περιπτώσεις δίνει μια καλή προσέγγιση της κατανομής.
5. Διατηρείται ή απορρίπτεται η μηδενική υπόθεση με βάση το αν η θεωρητική κατανομή υπερβαίνει την κρίσιμη τιμή του. Αν η στατιστική δοκιμή υπερβαίνει την κρίσιμη τιμή του χ^2 η μηδενική υπόθεση απορρίπτεται και η εναλλακτική γίνεται δεκτή. Αν η στατιστική δοκιμή πέσει κάτω από το όριο τότε διατηρείται η μηδενική υπόθεση, αν και δεν γίνεται απαραίτητα αποδεκτή.

3.2.4 Τεστ χ^2 καλής προσαρμογής

Το τεστ-χ^2 καλής προσαρμογής ελέγχει κατά πόσο μια γνωστή ή υποθετική κατανομή των δεδομένων, δηλαδή ένα μοντέλο, ταιριάζει με το σύνολο των πραγματικών παρατηρήσεων. Οι μετρήσεις καλής προσαρμογής συνοψίζουν την απόκλιση μεταξύ των πραγματικών τιμών και των αναμενόμενων τιμών που προκύπτουν από το μοντέλο. Χρειάζεται να βρεθούν οι βαθμοί ελευθερίας, οι αναμενόμενες συχνότητες, το στατιστικό και το ρ-value που σχετίζονται με το τεστ. Οι βαθμοί ελευθερίας είναι κ - 1, όπου κ τα επίπεδα ή οι ομάδες των κατηγορικών μεταβλητών. $E_i = n \cdot p_i$ είναι η αναμενόμενη συχνότητα για το i επίπεδο της κατηγορικής μεταβλητής, n το μέγεθος του δείγματος συνολικά και p_i το πληθυσμιακό ποσοστό των παρατηρήσεων στο i επίπεδο. Η τιμή του X^2 τεστ καταλληλότητας υπολογίζεται από τον τύπο $X^2 = \frac{(O_i - E_i)^2}{E_i}$ όπου, O_i η παρατηρούμενη συχνότητα για το i επίπεδο της κατηγορικής μεταβλητής και E_i η αναμενόμενη συχνότητα για το i επίπεδο της κατηγορικής μεταβλητής. Η υπόθεση του χ^2 τεστ καταλληλότητας δηλώνεται:

H_0 : Δεδομένα ακολουθούν μια καθορισμένη κατανομή.
H_1 : Δεδομένα δεν ακολουθούν την καθορισμένη κατανομή.
Η απόρριψη σημαίνει ότι το δείγμα είναι σημαντικά διαφορετικό από τον πληθυσμό
(Greenwood & Nikulin, 1996).

3.2.5 Τεστ χ^2 ανεξαρτησίας

Το τεστ χ^2 ανεξαρτησίας αξιολογεί αν κατηγορικές μεταβλητές συσχετίζονται μεταξύ τους στον ίδιο πληθυσμό. Σε δεδομένα ανάλυσης απαιτείται να βρεθούν οι βαθμοί ελευθερίας, οι αναμενόμενες συχνότητες, το στατιστικό και το p-value σχετικά με το τεστ. Μια παρατήρηση αποτελείται από τις τιμές δύο αποτελεσμάτων και η μηδενική υπόθεση είναι ότι η εμφάνιση αυτών των αποτελεσμάτων είναι στατιστικά ανεξάρτητη. Κάθε παρατήρηση εκχωρείται σε ένα κελί μιας δισδιάστατης διάταξης κελιών, που ονομάζεται πίνακας έκτακτης ανάγκης, σύμφωνα με τις τιμές των δύο αποτελεσμάτων. Οι βαθμοί ελευθερίας είναι ίσοι με $df = (r-1)\cdot(c-1)$, όπου r είναι οι γραμμές του πίνακα έκτακτης ανάγκης που αντιστοιχούν στον αριθμό των κατηγοριών σε μια μεταβλητή και c είναι οι στήλες που αντιστοιχούν στον αριθμό των κατηγοριών στη δεύτερη μεταβλητή (Bock & Velleman & De Veaux, 2007). Η αναμενόμενη συχνότητα για το κελί του επιπέδου r της μεταβλητής A και επιπέδου c της μεταβλητής B είναι $E_{r,c} = \frac{n_r \cdot n_c}{n}$, όπου

n_r : είναι το συνολικό δείγμα παρατηρήσεις στο επίπεδο r της μεταβλητής A.

n_c : ο συνολικός αριθμός του δείγματος στο επίπεδο c της μεταβλητής B.

n : συνολικό μέγεθος δείγματος.

Η τιμή της στατιστική δοκιμής είναι $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$,

όπου O_{ij} η παρατηρούμενη συχνότητα στο επίπεδο i της μεταβλητής A και στο επίπεδο j της μεταβλητής B. E_{ij} η αναμενόμενη συχνότητα στο επίπεδο i της μεταβλητής A και επίπεδο j της μεταβλητής B.

Υποθέσεις

H_0 : Συσχετισμένες μεταβλητές είναι ανεξάρτητες

H_1 : Σημαντικές μεταβλητές είναι συσχετισμένες

3.3 Παράδειγμα τεστ καλής προσαρμογής

Ένα χαρακτηριστικό παράδειγμα καλής προσαρμογής χ τετράγωνο είναι αυτό με το ζάρι. Για κάποιο λόγο υπάρχει η υποψία ότι ένα ζάρι δεν είναι δίκαιο και πρέπει να ελεγχθεί αυτή η υπόθεση. Θα συγκριθούν τα παρατηρούμενα δεδομένα από το πείραμα με τα αναμενόμενα δεδομένα από μια κατανομή. Έστω ότι σε 210 ρίψεις του ζαριού τα αποτελέσματα ήταν τα εξής: 1 ήρθε 23 φορές, 2 ήρθε 25 φορές, ..., 6 ήρθε 46 φορές. Ωστόσο, στη ρίψη ενός ζαριού τα δυνατά αποτελέσματα είναι ισοπίθανα με πιθανότητα P(X=1) = P(X=2) = P(X=3) = P(X=4)= P(X=5) = P(X=6) = $\frac{1}{6}$. Άρα σε 210 ρίψεις και σύμφωνα με το ισοπίθανο μοντέλο, που ισχύει από τις πιθανότητες, οι αναμενόμενες φορές να έρθει το 1 είναι $210\cdot\frac{1}{6}$ = 35 φορές, το ίδιο και το 2, το ίδιο και για τα άλλα αποτελέσματα. Το ισοπίθανο μοντέλο δίνει μια ομοιόμορφη κατανομή για τα αποτελέσματα του ζαριού. Έστω ότι η μηδενική υπόθεση H_0 είναι ότι το ζάρι είναι δίκαιο, δηλαδή ακολουθεί το ομοιόμορφο μοντέλο και ο αναμενόμενος αριθμός εμφάνισης για κάθε αριθμό είναι $210\cdot\frac{1}{6}$ = 35 .

Πίνακας 6: Παρατητούμενα και αναμενόμενα αποτελέσματα ρίψης ζαριού

Αποτέλεσμα ρίψης	1	2	3	4	5	6	Σύνολο

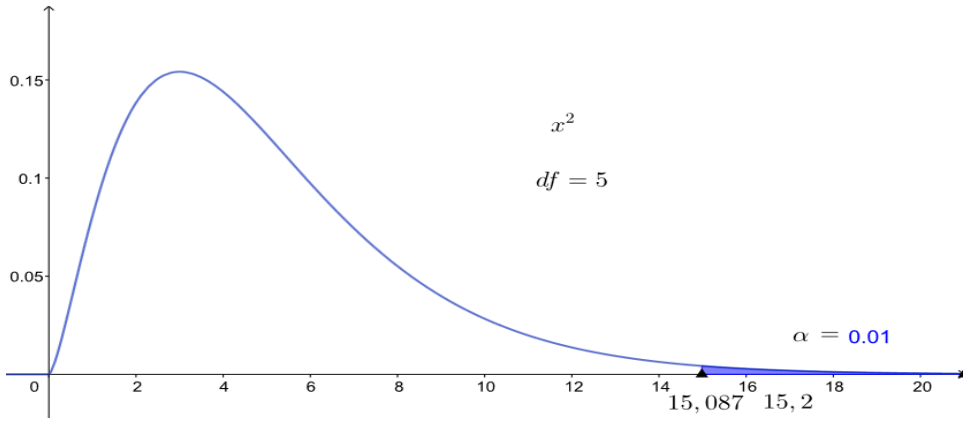
Παρατηρούμενα αποτελέσματα	20	25	42	41	36	46	210
Αναμενόμενα αποτελέσματα	35	35	35	35	35	35	210

Το τεστ υπόθεσης χ-τετράγωνο, όπως άλλωστε όλα τα τεστ υπόθεσης, έχει τα εξής βήματα:

1. Δηλώνεται η μηδενική υπόθεση
2. Επιλέγεται επίπεδο σημαντικότητας α
3. Υπολογίζονται οι κρίσιμες τιμές
4. Υπολογίζεται το στατιστικό
5. Αποτελέσματα

Βήμα 1. Η μηδενική υπόθεση H_0 είναι αυτή που επί του παρόντος πιστεύεται ότι είναι αληθής. Έστω ότι η μηδενική υπόθεση H_0 είναι ότι το ζάρι είναι δίκαιο. Δηλαδή οι παρατηρούμενες τιμές είναι κοντά ή ταιριάζουν στις αναμενόμενες τιμές. Ισοδύναμα H_0 δηλώνει ότι η ομοιόμορφη κατανομή ταιριάζει με τα δεδομένα. Η εναλλακτική υπόθεση H_1 δηλώνει ότι το ζάρι δεν είναι δίκαιο και ότι οι παρατηρούμενες τιμές δεν είναι κοντά ή δεν προσαρμόζονται με τις αναμενόμενες τιμές.

Βήμα 2. Επιλέγεται επίπεδο σημαντικότητας α. Το επίπεδο σημαντικότητας είναι η περιοχή στην ουρά. Στο σχήμα φαίνεται η κατανομή χ^2.



Γράφημα 7: Κατανομή χ^2 με 5 βαθμούς ελευθερίας

Αν δεν δίνεται το επίπεδο σημαντικότητας α, τότε το επιλέγει ο ερευνητής. Συνήθως είναι μεταξύ 0,01 και 0,05. Όσο μικρότερη είναι η τιμή του τόσο πιο σημαντικά θα είναι τα αποτελέσματα. Έστω ένα επίπεδο σημαντικότητας α = 0,01. Αυτό σημαίνει ότι η μπλε γραμμοσκιασμένη περιοχή στο σχήμα είνα ίση με 0,01. Αυτή η περιοχή λέγεται περιοχή απόρριψης και είναι σημαντική γιατί επιτρέπει να γίνει ένα συμπέρασμα στο τέλος του τεστ. Αν τα αποτελέσματα του τεστ πέσουν μέσα σε αυτή την περιοχή απόρριψης, τότε απορρίπτεται η μηδενική υπόθεση ότι το ζάρι είναι δίκαιο και γίνεται δεκτή η εναλλακτική.

Βήμα 3. Η κρίσιμη τιμή είναι το σημείο που χωρίζει την ουρά από το υπόλοιπο της καμπύλης. Θα είναι μια χ^2 τιμή αφού χρησιμοποείται το χ^2 τεστ. Η κρίσιμη τιμή υπολογίζεται από τον πίνακα χ^2, ανάλογα με τη σημαντικότητα και τους βαθμούς ελευθερίας. Η περιοχή στην ουρά είναι ίση με α = 0,01 και οι βαθμοί ελευθερία είναι πάντοτε κατά ένα λιγότεροι από τα πιθανά αποτελέσματα. Υπάρχουν 6 πιθανά αποτελέσματα άρα df = 6 – 1 = 5. Από τον πίνακα χ^2 για α = 0,01 και df = 5 η κρίσιμη τιμή είναι $\chi^2_{0,01}$ = 15,086.

Βήμα 4. Το στατιστικό του τεστ θα είναι χ^2 διότι γίνεται χ^2 τεστ. Ο τύπος σε αυτή την περίπτωση είναι $X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$, όπου O_i οι πραγματικές παρατηρήσεις και E_i οι αναμενόμενες παρατηρήσεις .

$$X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(20-35)^2}{35} + \frac{(25-35)^2}{35} + \frac{(42-35)^2}{35} + \frac{(41-35)^2}{35} + \frac{(36-35)^2}{35} + \frac{(46-35)^2}{35} =$$

$$\frac{225}{35} + \frac{100}{35} + \frac{49}{35} + \frac{36}{35} + \frac{1}{35} + \frac{121}{35} = \frac{532}{35} = 15{,}2$$

Επειδή το τεστ στατιστικό είναι μεγαλύτερο από την κρίσιμη τιμή 15,2 > 15,087, δηλαδή είναι δεξιά της κρίσιμης τιμής, βρίσκεται στην περιοχή απόρριψης. Αυτό σημαίνει ότι απορρίπτεται η μηδενική υπόθεση ότι το ζάρι είναι δίκαιο και γίνεται δεκτή η εναλλακτική υπόθεση ότι το ζάρι είναι μεροληπτικό.

Κεφάλαιο 4 Γενικευμένα γραμμικά μοντέλα

Η ευρεία χρήση υπολογιστών, στο πλαίσιο της στατιστικής ανάλυσης δεδομένων, βοήθησε στην ανάπτυξη νέων βελτιωμένων μεθόδων και μοντέλων. Τα γενικευμένα γραμμικά μοντέλα, τα οποία αποτελούν την συνέχεια των γραμμικών μοντέλων, μπορούν να χρησιμοποιηθούν σε πιο σύνθετες εφαρμογές.

4.1 Σύνθεση του μοντέλου

Η διαδικασία μοντελοποίησης περιέχει τα εξής τέσσερα βήματα:

1. Το μοντέλο ορίζεται από την εξίσωση που συνδέει την μεταβλητή απόκρισης και τις επεξηγηματικές μεταβλητές και την κατανομή πιθανότητας της μεταβλητής απόκρισης.
2. Εκτίμηση παραμέτρων του μοντέλου.
3. Έλεγχος για το πόσα καλά προσαρμόζεται το μοντέλο στα πραγματικά δεδομένα.
4. Εξαγωγή συμπερασμάτων, όπως έλεγχος υποθέσεων και υπολογισμός διαστημάτων εμπιστοσύνης για τις παραμέτρους του μοντέλου και ερμηνεία των αποτελεσμάτων (Dobson, 2002).

Συνήθως ακολουθείται ο παρακάτω συμβολισμός.

Πίνακας 7: Συμβολισμός στην στατιστική

Διάνυσμα τυχαίων μεταβλητών	$[Y_1, .., Y_n]^\tau$
Διάνυσμα παρατηρήσεων	$[y_1, .., y_n]^\tau$
Ο εκθέτης τ χρησιμοποιείται για ένα πίνακα όταν η στήλη γράφεται ως γραμμή	$[y_1, .., y_n]^\tau$
Εκτιμητής β ή εκτίμηση β	$\hat{\beta}$ ή b
Μέσος του y	$\bar{y} = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i$
Η συνάρτηση πυκνότητας πιθανότητας μιας συνεχούς τυχαίας μεταβλητής Y ή η συνάρτηση μάζας πιθανότητας αν Y είναι	f(y; θ)

διακριτή αναφέρεται απλά ως κατανομή πιθανότητας, με παράμετρο θ	
Η αναμενόμενη τιμή τυχαίας μεταβλητής Y	μ = E(Y)
Διακύμανση τυχαίας μεταβλητής Y	Var(Y)
Γραμμικός συνδυσαμός $Y_1, .., Y_n$	$W = \alpha_1 \cdot Y_1 + \ldots + \alpha_n \cdot Y_n$
Αναμενόμενη τιμή $W = \alpha_1 \cdot Y_1 + \ldots + \alpha_n \cdot Y_n$	$E(W) = \alpha_1 \cdot \mu_1 + \ldots + \alpha_n \cdot \mu_n$

4.1.1 Κατανομή πιθανότητας- Κανονική κατανομή

Η σύνθεση του μοντέλου γίνεται ανάλογα με το πως αποκτήθηκαν τα δεδομένα και το σχέδιο της έρευνας. Το μοντέλο έχει δύο συστατικά, την κατανομή πιθανοτήτων δηλαδή το σχήμα της κατανομής των δεδομένων και την εξίσωση που συνδέει την αναμενόμενη τιμή Y με τις επεξηγηματικές μεταβλητές. Για παράδειγμα μπορεί η κατανομή να είναι κανονική $Y \sim N(\mu, \sigma^2)$ και η εξίσωση $E(Y) = \alpha + \beta \cdot x$ ή $\ln E(Y) = \beta_0 + \beta_1 \cdot x$.

Ο συμβολισμός και η κωδικοποίηση για επεξηγηματικές μεταβλητές έχει ως εξής. Στα περισσότερα μοντέλα η εξίσωση σύνδεσης κάθε μεταβλητής απόκρισης Y και ενός συνόλου επεξηγηματικών μεταβλητών $x_1, x_2, \ldots\ldots, x_m$ έχει τύπο $g(E(Y)) = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_m \cdot x_m$. Για αποκρίσεις $Y_1, .., Y_N$ αυτό μπορεί να γραφεί ως $g(E(y)) = X \cdot \beta$ όπου $y = \begin{bmatrix} Y_1 \\ \ldots \\ Y_N \end{bmatrix}$ ένα διάνυσμα αποκρίσεων, $g(E(y)) = \begin{bmatrix} g(E(Y_1)) \\ \ldots \\ g(E(Y_N)) \end{bmatrix}$ δηλώνει ένα διάνυσμα συναρτήσεων των όρων $E(Y_i)$ με την ίδια g για κάθε στοιχείο, $\beta = \begin{bmatrix} \beta_1 \\ \ldots \\ \beta_\rho \end{bmatrix}$ είναι ένα διάνυσμα παραμέτρων και X είναι ένας πίνακας του οποίου τα στοιχεία είναι σταθερά και αντιπροσωπεύουν επίπεδα από κατηγορικές επεξηγηματικές μεταβλητές ή μετρήσιμες τιμές συνεχών επεξηγηματικών μεταβλητών. Για μια συνεχή επεξηγηματική μεταβλητή x, το μοντέλο περιέχει ένα αντίστοιχο όρο β_x, όπου η παράμετρος β_x αντιπροσωπεύει τη μεταβολή στην απόκριση για μεταβολή μιας μονάδας του x. Αν υπάρχουν ρ παράμετροι στο μοντέλο και N παρατηρήσεις, τότε y είναι ένας πίνακας Nx1 τυχαίο διάνυσμα, β είναι ένα ρx1 διάνυσμα παραμέτρων και X είναι ένας Nxp πίνακας γνωστών σταθερών. X συχνά λέγεται πίνακας σχεδίασης και $X \cdot \beta$ είναι η γραμμική συνιστώσα του μοντέλου.

Η επιλογή της κατάλληλης κατανομής γίνεται με διερεύνηση. Σε αυτή την περίπτωση χρησιμοποιούνται πίνακες συχνοτήτων, διάγραμματα κουκκίδων, ιστόγραμματα και άλλες γραφικές μεθόδοι. Πρόσθετα αναζητείται η συσχέτιση μεταξύ των μεταβλητών. Για συνεχείς μεταβλητές χρησιμοποιείται η γραφική παράσταση διασποράς, για να φανεί αν οι μεταβλητές συνδέονται με γραμμική ή μη γραμμική σχέση, για κατηγορικές μεταβλητές πίνακας σταυρού και για συνεχείς κλίμακας ομαδοποιημένες μετρήσεις τα box plot.

Πολλά στατιστικά τεστ χρησιμοποιούν την κανονική κατανομή για συνεχή δεδομένα που έχουν συμμετρική κατανομή. Αυτό συμβαίνει είτε απευθείας διότι σε πολλά φυσικά φαινόμενα οι τυχαίες μεταβλητές ακολουθούν την κανονική κατανομή, είτε έμμεσα από το κεντρικό οριακό θεώρημα για μεγάλα δείγματα. Αν η τυχαία μεταβλητή Y έχει κανονική κατανομή με μέσο μ και διακύμανση σ^2, η συνάρτηση πυκνότητας πιθανότητας είναι $f(y;\ \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp[-\frac{1}{2} \cdot (\frac{y-\mu}{\sigma^2})^2]$. Η κανονική κατανομή συμβολίζεται με $Y \sim N(\mu, \sigma^2)$. Αν $\mu=0$ και $\sigma^2=1$ τότε λέγεται τυπική κανονική κατανομή $Y \sim N(0,1)$. Έστω $Y_1, .., Y_n$ συμβολίζουν κανονικά κατανεμημένες τυχαίες μεταβλητές με $Y_i \sim N(\mu_i, \sigma_i^2)$ για i=1,.., n. Τότε η οριακή κατανομή των Y_i είναι η πολυμεταβλητή κανονική κατανομή με μέσο διάνυσμα $\mu= [\mu_1, \ldots, \mu_n]^\tau$.

Πολλές κατανομές παράγονται από την κανονική κατανομή. Κατανομές με αφετηρία την κανονική κατανομή είναι η χ^2, η F και η t. Αν $Z_1, .., Z_n$ ανεξάρτητες μεταβλητές που ακολουθούν την τυπική κανονική κατανομή $Zi \sim N(0, 1)$, i = 1,2,..., n, τότε η τυχαία μεταβλητή $X= Z_1^2 +\ldots+ Z_n^2$ ακολουθεί την κατανομή χ^2 με n βαθμούς ελευθερίας. Επίσης αν Z μια τυχαία μεταβλητή η οποία ακολουθεί την τυποποιημένη κανονική κατανομή, δηλαδή $Z \sim N(0, 1)$ και S_n μια τυχαία ανεξάρτητη μεταβλητή από την Z η οποία ακολουθεί την κατανομή χ_n^2 με n βαθμούς ελευθερίας, τότε η κατανομή της τυχαίας μεταβλητής $T = \frac{Z}{\sqrt{\frac{S_n}{n}}}$ λέγεται κατανομή t ή studen με n βαθμούς ελευθερίας. Τέλος αν $S_n \sim \chi_n^2$ και $S_m \sim \chi_m^2$ δύο ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την χ_n^2 και την χ_m^2 κατανομή αντίστοιχα τότε η τυχαία μεταβλητή $F = \frac{\frac{S_n}{n}}{\frac{S_m}{m}}$ ονομάζεται F κατανομή με n και m βαθμούς ελευθερίας (Dobson, 2002).

4.1.2 Εκτίμηση παραμέτρων

Η εκτίμηση της τιμής μιας άγνωστης παραμέτρου του υπό μελέτη πληθυσμού γίνεται διαισθητικά από ένα δείγμα του πληθυσμού. Αν $y_1, .., y_n$ οι παρατηρήσεις ενός τυχαίου δείγματος το ζητούμενο είναι ένας αριθμός που μπορεί να θεωρηθεί η τιμή της εκτίμησης. Εκτιμήτρια είναι μια τυχαία μεταβλητή που χρησιμοποιείται για να εκτιμήσει ένα

χαρακτηριστικό του πληθυσμού, δηλαδή για να εκτιμήσει την παράμετρο. Η τιμή της εκτιμήτριας λέγεται εκτίμηση. Πολλές φορές χρησιμοποιείται διαισθητικά ως εκτιμήτρια η αντίστοιχη συνάρτηση του δείγματος. Για παράδειγμα η εκτιμήτρια της μέσης τιμής μ ενός πληθυσμού είναι ο δειγματικός μέσος $\bar{y}$. Η σημειακή εκτίμηση δίνει μια μόνο τιμή ως εκτίμηση της παραμέτρου και οι μέθοδοι της σημειακής εκτίμησης είναι η μέθοδος μέγιστης πιθανοφάνειας, η μέθοδος ελαχίστων τετραγώνων και η μέθοδος των ροπών. Για διάστημα τιμών η εκτίμηση είναι διάστημα εμπιστοσύνης.

4.1.2.1 Μέθοδος μέγιστης πιθανοφάνειας

Εστω $y = [Y_1, .., Y_n]^\tau$ συμβολίζει ένα τυχαίο διάνυσμα και έστω η κοινή συνάρτηση πυκνότητας πιθανότητας των Y_i είναι f(y;θ) που εξαρτάται από την παράμετρο $\theta = [\theta_1, \ldots, \theta_\rho]^\tau$. Ορίζεται η συνάρτηση πιθανότητας L(θ;y) η οποία είναι αλγεβρικώς η ίδια με την κοινή συνάρτηση πυκνότητας πιθανότητας f(y;θ), ωστόσο η αλλαγή στον τύπο αντανακλά μια μετατόπιση έμφασης από τις τυχαίες μεταβλητές y με σταθερό θ, στις παραμέτρους θ με σταθερό y. Αφού το L ορίζεται ως προς το τυχαίο διάνυσμα y, είναι και αυτό μια τυχαία μεταβλητή. Η πιθανότητα L(θ; y) είναι συνήθως συνάρτηση της άγνωστης παραμέτρου θ. Για κάθε τιμή του θ η πιθανότητα του ενδεχομένου y αλλάζει. Η τιμή του θ για την οποία το ενδεχόμενο y έχει την μεγαλύτερη πιθανότητα να συμβεί λέγεται εκτιμήτρια του θ και συμβολίζεται με $\hat{\theta}$. Αρκεί να βρεθεί το θ που μεγιστοποιεί την συνάρτηση πιθανότητας, $L(\hat{\theta}; y) \geq L(\theta; y)$ για κάθε θ. Για να βρεθεί το μέγιστο της συνάρτησης f(y;θ) ή L(θ; y) αρκεί να βρεθεί το μέγιστο του φυσικού λογάριθμου l(θ) = logL(θ;y). Άρα η εκτίμηση μέγιστης πιθανότητας $\hat{\theta}$ της θ είναι αυτή που μεγιστοποιεί την L(θ;y) ή την l(θ) = logL(θ;y). Ο υπολογισμός της μέγιστης τιμής γίνεται με παραγώγιση και ακρότατα στη συνάρτηση l(θ) (Dobson, 2002).

Παράδειγμα Poisson κατανομή

Για να εκτιμηθεί ο αναμενόμενος αριθμός γκολ που βάζει ο Ολυμπιακός στα εντός έδρας παιχνίδια του αρκεί να εκτιμηθεί η παράμετρο θ της κατανομής Poisson. Έστω ότι ο αριθμός των γκολ εντός έδρας σε n αγώνες είναι $Y_1, .., Y_n$ και έστω ότι τα γκολ μπαίνουν τυχαία σε κάθε παιχνίδι. Έστω $Y_1, .., Y_n$ ανεξάρτητες τυχαίες μεταβλητές κάθε μια με Poisson κατανομή $f(y_i ; \theta) = \frac{\theta^{y_i} \cdot e^{-\theta}}{y_i!}$, $y_i = 0,1,\ldots$ με την ίδια παράμετρο θ και άρα η πιθανότητα να βάλει ο Ολυμπιακός y_i γκολ σε ένα αγώνα i είναι Poisson με παράμετρο θ, $P(Y_i=y_i) = f(y_i; \theta) = e^{-\theta} \cdot \frac{\theta^{y_i}}{y_i!}$, $y_i = 0, 1,\ldots, n$. Τα παιχνίδια είναι ανεξάρτητα μεταξύ τους άρα η πιθανότητα των n μετρήσεων $y_1,.., y_n$

ή η κοινή κατανομή είναι

$$f(y_1,..,y_n,\theta) = P(Y_1=y_1,..,Y_n=y_n) = \prod_{i=1}^{n} e^{-\theta} \cdot \frac{\theta^{y_i}}{y_i!} = \frac{e^{-n\theta}\cdot\theta^{\Sigma y_i}}{y_1!\cdots y_n!} = L(\theta)$$

Άρα $l(\theta) = \log L(\theta) = \log \frac{e^{-n\theta}\cdot\theta^{\Sigma y_i}}{y_1!\cdots y_n!} = -n\theta + \Sigma y_i \cdot \log\theta - \Sigma \log(y_i!)$. Με παραγώγιση της συνάρτησης $l(\theta)$ ως προς θ και μονοτονία, υπολογίζεται ότι το μέγιστο είναι $\theta = \frac{\Sigma y_i}{n}$. Άρα $\hat{\theta} = \bar{y}$ είναι η εκτίμηση μέγιστης πιθανοφάνειας (Dobson, 2002).

4.1.2.2 Μέθοδος ελαχίστων τετραγώνων

Η μέθοδος αυτή βασίζεται στην εύρεση της εκτιμήτριας με ελαχιστοποίηση του αθροίσματος τετραγώνων. Για παράδειγμα έστω ότι πρέπει να εκτιμηθεί η μέση τιμή μ ενός πληθυσμού με βάση ένα τυχαίο δείγμα $Y_1,..,Y_n$. Τότε η εκτίμηση του μ είναι το ελάχιστο της παράστασης $S = \sum_{i=1}^{n}(Y_i - \mu)^2$ ως προς μ. Με παραγώγιση υπολογίζεται το ελάχιστο $\mu = \frac{Y_1+\cdots+Y_n}{n}$, άρα $\hat{\mu} = \bar{y}$.

Γενικά για την εκτίμηση μιας παραμέτρου θ ή ενός διανύσματος παραμέτρων $\theta = [\theta_1,\ldots,\theta_\rho]^\tau$ με βάση ένα σύνολο n παρατηρήσεων $Y_1,..,Y_n$ με αναμενόμενες τιμές $\mu_1,...,\mu_n$ αντίστοιχα οι οποίες συνδέονται με την υπό εκτίμηση παράμετρο θ μέσω των μέσων τιμών $E(Y_i) = \mu_i(\theta)$, τότε η εκτιμήτρια $\hat{\theta}$ της παραμέτρου θ είναι εκείνη που ελαχιστοποιεί το άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των τιμών Y και των αντίστοιχων αναμενόμενων τιμών. Δηλαδή αυτή που ελαχιστοποιεί την παράσταση $S = \sum_{i=1}^{n}(Y_i - E(Y_i))^2 = \sum_{i=1}^{n}(Y_i - \mu_i(\theta))^2$. Το $\hat{\theta}$ υπολογίζεται με παραγώγιση του S σε σχέση με κάθε στοιχείο θ_j του θ και λύνοντας το σύστημα των εξισώσεων $\frac{dS}{d\theta_j} = 0$, $j = 1,2,\ldots,p$. Σε πολλές καταστάσεις η μέθοδοι μέγιστης πιθανότητας και ελαχίστων τετραγώνων δεν παρουσιάζουν διαφορές, ωστόσο στην μέθοδο ελαχίστων τετραγώνων δεν χρειάζεται να προσδιοριστεί η κοινή κατανομή πιθανότητας των Y_i (Dobson, 2002).

4.2 Έλεγχος μοντέλου και καταλοίπων

Τα κατάλοιπα είναι σημαντικά διαγνωστικά εργαλεία για τον έλεγχο των υποθέσεων του μοντέλου παλινδρόμησης. Τα μοντέλα γραμμικής παλινδρόμησης προϋποθέτουν ότι τα κατάλοιπα είναι ανεξάρτητα, έχουν κανονική κατανομή και σταθερή διασπορά. Ο έλεγχος κανονικότητας γίνεται με το γράφημα της κανονικής πιθανότητας, όπου σχεδιάζονται τα κατάλοιπα με τις αναμενόμενες τιμές τους, από την κανονική κατανομή, και τα σημεία στο γράφημα πρέπει να βρίσκονται πάνω ή να είναι κοντά στην ευθεία γραμμή που αντιπροσωπεύει

την κανονικότητα. Αλλιώς η υπόθεση της κανονικότητας δεν είναι επαρκής. Επίσης από τα κατάλοιπα ελέγχεται η γραμμικότητα των μεταβλητών και η μεταβολή της διακύμανσης. Η υπόθεση σταθερής διακύμανσης λέγεται ομοσκεδαστικότητα (Dobson, 2002).

Τέλος τα τυποποιημένα κατάλοιπα σχεδιάζονται για κάθε μια από τις επεξηγηματικές μεταβλητές του προσαρμοσμένου μοντέλου. Αν το μοντέλο περιγράφει επαρκώς την επίδραση της μεταβλητής τότε δεν πρέπει να υπάρχει εμφανές μοτίβο στο γράφημα. Αν το μοντέλο δεν είναι επαρκές τότε τα σημεία μπορούν να εμφανίζουν καμπυλότητα ή κάποιο άλλο μοτίβο το οποίο δηλώνει ότι πρέπει να γίνουν αλλαγές στο μοντέλο. Παρόμοια γίνεται ο έλεγχος για το αν πρέπει να περιληφθούν πρόσθετες μεταβλητές στο μοντέλο ή αν πρέπει να αφαιρεθούν.

Έστω ότι τα κατάλοιπα ακολουθούν την κανομική κατανομή και η μεταβλητή απόκρισης Y μοντελοποιείται ως $E(Y_i) = \mu_i$, $Y_i \sim N(\mu_i , \sigma^2)$. Οι εκτιμώμενες τιμές του μ_i είναι $\hat{\mu}_i$. Τα κατάλοιπα τότε μπορούν να οριστούν ως $y_i - \hat{\mu}_i$ όπου y_i τα παρατηρούμενα σημεία, ενώ τα τυποποιημένα κατάλοιπα ως $r_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}}$ όπου $\hat{\sigma}$ είναι η εκτίμηση της άγνωστης παραμέτρου σ. Αυτά τα τυποποιημένα κατάλοιπα είναι ελαφρώς συσχετισμένα διότι όλα εξαρτώνται από τις εκτιμήσεις $\hat{\mu}_i$ και $\hat{\sigma}$ που έχουν υπολογιστεί από τις παρατηρήσεις. Επίσης δεν είναι ακριβώς κανονικά κατανεμημένα διότι το σ έχει εκτιμηθεί ως $\hat{\sigma}$. Παρ'όλα αυτά είναι προσεγγιστικά κανονικά κατανεμημένα και η επάρκεια της προσέγγισης μπορεί να ελεγχθεί χρησιμοποιώντας γραφικές μεθόδους. Οι παράμετροι μ_i είναι συναρτήσεις των επεξηγηματικών μεταβλητών. Αν το μοντέλο περιγράφει καλά τη σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών αυτό θα φανεί ή θα εξηγηθεί από τα $\hat{\mu}_i$. Η διαφορά $y_i - \hat{\mu}_i$ θα πρέπει να είναι μικρή, κάτι το οποίο φαίνεται και γραφικά. Επίσης το άθροισμα των τετραγώνων των καταλοίπων $\Sigma(y_i - \hat{\mu}i)^2$ παρέχει ένα συνολικό στατιστικό για αξιολόγηση της επάρκειας του μοντέλου.

Για άλλες κατανομές, εκτός από την κανονική, χρησιμοποιείται μια ποικιλία ορισμών των τυποποιημένων καταλοίπων. Μερικοί από αυτούς είναι μετατροπές των όρων $y_i - \hat{\mu}_i$ σχεδιασμένοι να βελτιώσουν την κανονικότητα ή την ανεξαρτησία. Στην περίπτωση που τα κατάλοιπα είναι από Poisson μοντέλο τότε $E(Y_i) = \theta_i$, $Y_i \sim$ Poisson (θ_i) και τα τυποποιημένα κατάλοιπα έχουν κατά προσέγγιση τύπο $r_i = \frac{y_i - \hat{\theta}i}{\sqrt{\hat{\theta}i}}$.

4.2.1 Παράδειγμα γκολ εντός έδρας και εκτός έδρας

Η διαδικασία να βρεθεί το κατάλληλο προσαρμοσμένο μοντέλο πολλές φορές περιλαμβάνει σύγκριση διαφορετικών μοντέλων κάτω από διαφορετικές υποθέσεις. Τότε το

ερώτημα αν τα δεδομένα υποστηρίζουν μια συγκεκριμένη υπόθεση μπορεί να διατυπωθεί ως προς την επάρκεια της προσαρμογής του αντίστοιχου μοντέλου που είναι σχετικό με άλλα πιο περίπλοκα μοντέλα.

Δεδομένα από το ελληνικό πρωτάθλημα τη σεζόν 1994-95 περιέχουν τα γκολ εντός έδρας και εκτός έδρας που σημείωσε η ΑΕΚ σε 32 παιχνίδια, 16 εντός και 16 εκτός έδρας, και δείχνουν ότι τα γκολ εντός έδρας τείνουν να είναι πιο πολλά. Δεν είναι ξεκάθαρο αν αυτό συμβαίνει επειδή παίζει καλύτερα εντός έδρας ή εξαιτίας παραγόντων όπως η επίδραση της έδρας. Ο πίνακας δείχνει τους αριθμούς των γκολ που αναφέρονται σε αγώνες της ΑΕΚ τη σεζόν 1994-95 σε εντός έδρας αγώνες και σε εκτός έδρας αγώνες.

Πίνακας 8: Αριθμός γκολ της ΑΕΚ σε 16 εντός και σε 16 εκτός έδρας αγώνες

Εντός	0	1	2	2	2	4	2	3	4	5	5	3	1	4	0	0
n=16, μέσος = 2,4 , τυπική απόκλιση = 1,7 ,διακύμανση = 2,9																
Εκτός	2	1	1	1	0	3	1	1	0	0	0	2	3	1	1	2
n=16, μέσος =1,2 , τυπική απόκλιση = 0,98 ,διακύμανση=0,96																

Το ερώτημα είναι αν η ομάδα της ΑΕΚ έχει την ίδια ικανότητα να σκοράρει στα εντός όσο και στα εκτός έδρας παιχνίδια, όπως αποδεικνύεται από τον αριθμό των γκολ εντός και εκτός. Η Poisson κατανομή παρέχει ένα αληθοφανές τρόπο μοντελοποίησης αυτών των δεδομένων καθώς πρόκειται για καταμέτρηση και σε κάθε περίπτωση, εντός ή εκτός έδρας, ο δειγματικός μέσος και η διακύμανση είναι προσεγγιστικά ίσα.
Έστω Y_{jk} τυχαία μεταβλητή που αντιπροσωπεύει των αριθμό των γκολ για το k παιχνίδι στην j περίπτωση, όπου j=1 αν ο αγώνας είναι εντός και j=2 αν ο αγώνας είναι εκτός έδρας και

k =1, 2, ..., K_j με K_1 =16 και K_2 =16.

Έστω ότι τα Y_{jk} είναι όλα ανεξάρτητα και έχουν την Poisson κατανομή με παράμετρο θ_j που αντιπροσωπεύει τον αναμενόμενο αριθμό γκολ. Το ερώτημα ενδιαφέροντος μπορεί να διατυπωθεί ως ένα τεστ της μηδενικής υπόθεσης H_0 εναντίον της εναλλακτικής H_1.

H_0: $\theta_1 = \theta_2 = \theta$

H_1: $\theta_1 \neq \theta_2$

Δηλαδή η μηδενική υπόθεση λέει ότι ο αναμενόμενος αριθμός γκολ εντός και ο αναμενόμενος αριθμός γκολ εκτός είναι ίσοι.

Για να γίνει έλεγχος της H_0 υπόθεσης προσαρμόζονται 2 μοντέλα:

1ο μοντέλο: Θεωρεί ότι H_0 είναι αληθής δηλαδή $E(Y_{jk}) = \theta$, $Y_{jk} \sim Poisson(\theta)$

2ο μοντέλο: Θεωρεί ότι δεν είναι αληθής η H_0, δηλαδή $E(Y_{jk}) = \theta_j$, $Y_{jk} \sim Poisson(\theta_j)$ όπου j =1 ή 2.

Για να ελεγχθεί η H_0 αρκεί να συγκριθούν τα μοντέλα στο πόσο καλά προσαρμόζονται στα δεδομένα. Αν είναι εξίσου καλά τότε δεν υπάρχει λόγος απόρριψης της H_0. Ωστόσο αν το μοντέλο 2 είναι ξεκάθαρα καλύτερο τότε H_0 θα απορριφθεί υπέρ την H_1.

Αν H_0 είναι αληθής τότε η log πιθανότητα συνάρτηση του Y_{jk} είναι $l(\theta;y) = \sum_{j=1}^{J} \sum_{k=1}^{K_j} (y_{jk} \cdot \log\theta - \theta - \log y_{jk}!)$ όπου J = 2 σε αυτή την περίπτωση. Η εκτίμηση μέγιστης πιθανοφάνειας είναι

$\hat{\theta} = \sum \sum \frac{y_{jk}}{N} = \frac{57}{32} = 1{,}78$, όπου

$N = \sum_j K_j = K_1 + K_2 = 16+16 = 32$.

Αν η H_1 είναι αληθής τότε η log συνάρτηση πιθανότητας είναι

$l(\theta_1, \theta_2;y) = \sum_{k=1}^{K_1} (y_{1k} \cdot \log\theta_1 - \theta_1 - \log y_{1k}!) + \sum_{k=1}^{K_2} (y_{2k} \cdot \log\theta_2 - \theta_2 - \log y_{2k}!)$

Η εκτίμηση μέγιστης πιθανότητας είναι $\hat{\theta}_j = \sum_k \frac{y_{jk}}{K_j}$ για j=1 ή 2. Σε αυτή την περίπτωση $\hat{\theta}_1 = \frac{38}{16} = 2{,}375$ και $\hat{\theta}_2 = \frac{19}{16} = 1{,}1875$.

Η μέγιστη τιμή της log συνάρτησης πιθανότητας $l(\hat{\theta}_1 , \hat{\theta}_2 ;y)$ θα είναι πάντα μεγαλύτερη ή ίση από την log συνάρτηση της $l(\hat{\theta};y)$ διότι μια επιπλέον παράμετρος έχει προστεθεί. Για να αποφασιστεί αν η διαφορά είναι στατιστικά σημαντική εξετάζεται η αναλογία πιθανότητας $\lambda = \frac{L(\hat{\theta}1 , \hat{\theta}2 ; y)}{L(\hat{\theta};y)}$ η οποία παρέχει ένα τρόπο αξιολόγησης ή καλής προσαρμογής των μοντέλων (ratio test). Αποδεικνύεται ότι $2\log\lambda = 2(l(\hat{\theta}_1 , \hat{\theta}_2 ; y) - l(\hat{\theta};y))$ ακολουθεί χ^2 κατανομή, και λέγεται απόκλιση (deviance). Αν $2\log\lambda \geq \chi^2_{1-\alpha}(r)$ όπου r η διαφορά των παραμέτρων των μοντέλων, τότε απορρίπτεται η μηδενική υπόθεση

H_0: $\theta_1 = \theta_2 = \theta$.

$l(\hat{\theta}_1 , \hat{\theta}_2 ; y) = -51{,}9932$ και $l(\hat{\theta};y) = -55{,}2213$

$2(l(\hat{\theta}_1 , \hat{\theta}_2 ; y) - l(\hat{\theta};y)) = 2 \cdot 3{,}228094 = 6{,}456189$

ενώ $\chi^2_{0,05}(1) = 3,841$

Άρα απορρίπτεται η μηδενική υπόθεση.

Αν Y~ poisson (θ) τότε E(Y) = Var(Y) = θ. Η εκτίμηση $\hat{\theta}$ του E(Y) λέγεται προσαρμοσμένη τιμή του Y. Η διαφορά Y- $\hat{\theta}$ λέγεται κατάλοιπο. Ένα κατάλοιπο είναι συνήθως τυποποιημένο με διαίρεση με το τυπικό σφάλμα του, ενώ για την Poisson κατανομή μια προσέγγιση τυποποιημένου κατάλοιπου είναι r = $\frac{Y-\hat{\theta}}{\sqrt{\hat{\theta}}}$.

Τα τυποποιημένα κατάλοιπα των μοντέλων 1 και 2 φαίνονται στον πίνακα.

Πίνακας 9: Τυποποιημένα κατάλοιπα των μοντέλων 1 και 2

Τιμή του Y	Συχνότητα	Τυποποιημένα κατάλοιπα από μοντέλο 1	Τυποποιημένα κατάλοιπα από μοντέλο 2
		$\hat{\theta} = 1,78$	$\hat{\theta}_1 = 2,38$
Εντός έδρας			
0	3	-1,334	-1,543
1	2	-0,585	-0,894
2	4	0,165	-0,246
3	2	0,9145	0,402
4	3	1,664	1,05
5	2	2,4138	1,698
			$\hat{\theta}_2 = 1,19$
Εκτός έδρας			
0	4	-1,334	-1,09
1	7	-0,585	-0,174

2	3	0,165	0,743
3	2	0,9145	1,66

Τα κατάλοιπα μπορούν να χρησιμοποιηθούν για την επάρκεια του μοντέλου. Για την Poisson κατανομή με ανεξάρτητες τυχαίες μεταβλητές Y_i, αν τα θ_i δεν είναι πολύ μικρά, τα τυποποιημένα κατάλοιπα $r_i = \frac{Yi - \hat{\theta}i}{\sqrt{\hat{\theta}i}}$ έχουν προσεγγιστικά την τυπική κανονική κατανομή N ~ (0,1), αν και συνήθως δεν είναι ανεξάρτητα. Διαισθητικά ισχύει ότι αν $r_i \sim N(0,1)$, τότε $r_i^2 \sim x^2(1)$ και άρα $\Sigma r_i^2 = \Sigma \frac{(Y_i - \hat{\theta}i)^2}{\hat{\theta}i} \sim x^2(m)$, όπου m ο αριθμός των παραμέτρων που εκτιμήθηκαν προκειμένου να υπολογιστούν οι προσαρμοσμένες τιμές $\hat{\theta}$. Η παραπάνω σχέση είναι το συνηθισμένο χ^2 τεστ καλής προσαρμογής για καταμέτρηση δεδομένων που συνήθως γράφεται $X^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} \sim x^2(m)$ όπου O_i είναι οι παρατηρούμενες συχνότητες και E_i είναι οι αντίστοιχες αναμενόμενες συχνότητες. Σε αυτή την περίπτωση $O_i = Y_i$, $E_i = \hat{\theta}_i$ και $\Sigma r_i^2 = X^2$.

Για το μοντέλο 1:

$\Sigma r_i^2 = 3\cdot(-1{,}334)^2 + 2\cdot(-0{,}585)^2 + \ldots\ldots + 2\cdot(0{,}9145)^2 = 39{,}06$

Αυτή η τιμή είναι μικρότερη από την κρίσιμη τιμή της κεντρικής χ^2 κατανομής με

m =16+16-1=31 βαθμούς ελευθερίας $\chi^2_{0,05}(31) = 44$

Όμοια για το μοντέλο 2:

$\Sigma r_i^2 = 3\cdot(-1{,}543)^2 + 2\cdot(-0{,}894)^2 + \ldots\ldots + 3\cdot(0{,}743)^2 + 2\cdot(1{,}66)^2 = 30{,}85$

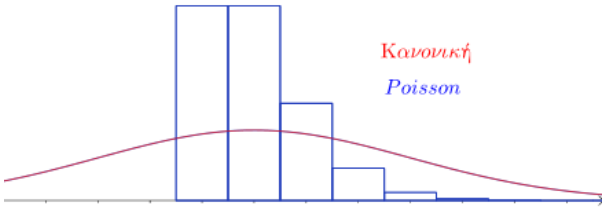
που συμφωνεί με την κεντρική χ^2 κατανομή με m = 32-2 = 30 βαθμούς ελευθερίας διότι

$\chi^2_{0,05}(30) = 43{,}7$.

Άρα κανένα από τα δύο μοντέλα δεν απορρίπτεται. Ωστόσο η διαφορά τιμών Σr_i^2 από το μοντέλο 1 και 2 είναι μεγάλη 39,06 - 30,85 = 8,21. Αυτό δείχνει ότι το μοντέλο 2 με μικρότερο Σr_i^2 μπορεί να περιγράψει τα δεδομένα πολύ καλύτερα από το μοντέλο 1. Αυτό σημαίνει ότι τα δεδομένα παρέχουν αποδείξεις που υποστηρίζουν ότι δεν ισχύει η μηδενική υπόθεση H_0: $\theta_1 = \theta_2$ (Dobson, 2002).

4.3 Γενικευμένα γραμμικά μοντέλα

Σε πολλές περιπτώσεις γίνεται η υπόθεση ότι η κατανομή είναι κανονική. Ωστόσο σε πολλά πραγματικά προβλήματα τα δεδομένα δεν είναι κανονικά κατανεμημένα.



Γράφημα 8: Διάγραμμα της κανονικής και το ιστόγραμμα της Poisson κατανομής

Στο παραπάνω σχήμα φαίνονται το διάγραμμα της κανονικής και το ιστόγραμμα της Poisson κατανομής τα οποία έχουν τους ίδιους μέσους μ=1. Είναι εμφανής η διαφορά όσον αφορά το πως κατανέμονται οι τιμές. Ο υπολογισμός πολλών στατιστικών όπως η συνάρτηση πυκνότητας πιθανότητας, τα διαστήματα εμπιστοσύνης, τα p-values, υποθέτοντας ότι η κατανομή είναι κανονική, ενώ στην πραγματικότητα είναι Poisson ή μια άλλη κατανομή, θα μπορούσε να οδηγήσει σε λάθη. Η κανονική κατανομή δεν ισχύει για όλες τις καταστάσεις. Για παράδειγμα μια πάθηση που οφείλεται στο τσιγάρο δεν είναι κανονική κατανομή, διότι υπάρχει μεγάλο μέρος του πληθυσμού που δεν καπνίζει καθόλου ή καπνίζει πολύ λίγο. Είναι σημαντικό να επιλεγεί μια κατανομή ανάλογα με το πως συμπεριφέρονται τα δεδομένα κάθε φορά (Quantitative Psychology, 2018).

Επίσης αν τα δεδομένα δεν είναι κανονικά κατανεμημένα δεν μπορεί να εφαρμοστεί το γραμμικό μοντέλο. Η γραμμική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της τιμής της συνεχούς μεταβλητής Y από τον γραμμικό συνδυασμό των επεξηγηματικών μεταβλητών X και βασικές προϋποθέσεις είναι ότι η μεταβλητή απόκρισης Y ακολουθεί κανονική κατανομή, η σχέση μεταξύ των μεταβλητών X και Y είναι γραμμική, η διακύμανση είναι σταθερή και το Y είναι συνεχής μεταβλητή. Όλοι αυτοί οι περιορισμοί είναι δεσμευτικοί στην γραμμική παλινδρόμηση και ταυτόχρονα είναι δύσκολο να ισχύουν για όλα τα δεδομένα. Ένα μοντέλο όπου έχει μεθόδους ανάλογες με τα γραμμικά μοντέλα, αλλά εφαρμόζεται και σε πιο γενικές καταστάσεις, είναι το γενικευμένο γραμμικό μοντέλο το οποίο αναπτύχθηκε από τους Nelder και Wedderburn το 1972. Το γενικευμένο γραμμικό μοντέλο χρησιμοποιείται σε πραγματικά δεδομένα, ενώ έχει πολλά πλεονεκτήματα σε σχέση με το γραμμικό μοντέλο. Για παράδειγμα οι μεταβλητές απόκρισης μπορεί να έχουν κατανομές άλλες από την κανονική κατανομή, δεν

απαιτείται τα δεδομένα να είναι γραμμικά, η διακύμανση μπορεί να μην είναι σταθερή, οι μεταβλητές απόκρισης μπορεί να είναι κατηγορικές.

Το γενικευμένο γραμμικό μοντέλο αποτελεί την συνέχεια του γραμμικού μοντέλου. Γενικεύοντας τη μορφή $E(Y_i) = \mu_i = x_i^{\tau} \cdot \beta$, $Y_i \sim N(\mu, \sigma^2)$ και κάποιες ιδιότητες της κανονικής κατανομής, αναπτύχθηκε μια ευρύτερη οικογένεια κατανομών που λέγεται εκθετική οικογένεια κατανομών. Ο αριθμητικός τρόπος υπολογισμού των παραμέτρων β στο γραμμικό μοντέλο $E(Y_i) = \mu_i = x_i^{\tau} \cdot \beta$, $Y_i \sim N(\mu, \sigma^2)$ επεκτάθηκε και σε καταστάσεις όπου η σχέση μεταξύ του μ_i και των παραμέτρων x_i δεν είναι γραμμική, αλλά αντίθετα υπάρχει μια συνάρτηση g τέτοια ώστε $g(\mu_i) = x_i^{\tau} \cdot \beta$. Η συνάρτηση g λέγεται συνάρτηση σύνδεσης και ο σκοπός της είναι να χαλαρώσει τις υποθέσεις που επιβάλλουν κάποιες κατανομές.

Πίνακας 10: Εξίσωση γραμμικού και γενικευμένου γραμμικού μοντέλου

Κανονική κατανομή στο γραμμικό μοντέλο	$\mu_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$
Εκθετική οικογένεια κατανομών στο γενικευμένο γραμμικό μοντέλο	$g(\mu_i) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$

Το ζητούμενο σε ένα μοντέλο είναι να υπολογιστούν οι συντελεστές των προβλεπτικών παραγόντων β_0, β_1, β_2. Στην εκθετική οικογένεια κατανομών αφού εκτιμηθούν τα β_i, θα εκτιμηθεί στη συνέχεια το $g(\mu_i)$, και τέλος το μ_i.
Η συνάρτηση σύνδεσης g είναι συνήθως μια γνωστή μαθηματική συνάρτηση. Επίσης οι αριθμητικοί υπολογισμοί γίνονται με τη βοήθεια στατιστικών προγραμμάτων στον υπολογιστή και το διάνυσμα β των συντελεστών παλινδρόμησης εκτιμάται από την μέγιστη πιθανότητα χρησιμοποιώντας αλγόριθμο με τα επαναληπτικά σταθμισμένα ελάχιστα τετράγωνα.

Το γενικευμένο γραμμικό μοντέλο αποτελείται από τρία συστατικά:
1. Το πρώτο συστατικό είναι το συστηματικό συστατικό $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$, δηλαδή η συνάρτηση που συνδέει τους προβλεπόμενους παράγοντες x_1, x_2 με το αποτέλεσμα, όπως συμβαίνει και στο γραμμικό μοντέλο.
2. Το δεύτερο συστατικό είναι μια μαθηματική συνάρτηση που στρέφει τη γραμμή της αρχικής συνάρτησης και λέγεται συνάρτηση σύνδεσης. Για παράδειγμα οι γνωστές συναρτήσεις $g(x) = x^2$, $g(x) = \sqrt{x}$, $g(x) = e^x$, $g(x) = \log x$ μπορεί να είναι συναρτήσεις σύνδεσης.

3. Το τρίτο συστατικό είναι η κατανομή που ακολουθούν οι μεταβλητές, η οποία πρέπει να ανήκει σε μια συγκεκριμένη οικογένεια κατανομών, που τις χαρακτηρίζει ένας κοινός τύπος εξίσωσης. Η οικογένεια αυτή λέγεται εκθετική οικογένεια κατανομών και περιλαμβάνει πολλές κατανομές όπως κανονική, Poisson, αρνητική διωνυμική, Γάμμα, Beta (Quantitative Psychology, 2018).

Γενικεύοντας, έστω ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, .., Y_N$ που έχουν μια κατανομή από την εκθετική οικογένεια κατανομών. Έστω $E(Y_i) = \mu_i$ όπου μ_i είναι κάποια συνάρτηση της παραμέτρου θ_i. Για το γενικευμένο γραμμικό μοντέλο υπάρχει ο μετασχηματισμός του μ_i ώστε $g(\mu_i) = \beta_1 \cdot x_{i1} + ... + \beta_p \cdot x_{ip}$, όπου $\beta_1, \ldots, \beta_p$ σύνολο συντελεστών των προβλεπτικών παραγόντων ή επεξηγηματικών μεταβλητών $x_{i1}, \ldots, x_{ip}$, ($p<N$), $x_i = \begin{bmatrix} x_{i1} \\ \cdots \\ x_{ip} \end{bmatrix}$. Η g είναι η συνάρτηση σύνδεσης.

Τα στοιχεία του γενικευμένου γραμμικού μοντέλου είναι : Η συστημική συνιστώσα, η λειτουργία συνδέσμου και το τυχαίο συστατικό.

Συστηματική συνιστώσα

Το συστηματικό συστατικό είναι το σύνολο παραμέτρων και επεξηγηματικών μεταβλητών

$$\beta_1 \cdot x_{i1} + ... + \beta_p \cdot x_{ip} = \sum_{j=1}^{p} \beta_j \cdot x_{ij} = [x_{i1}, \ldots, x_{ip}] \cdot \begin{bmatrix} \beta_1 \\ \cdots \\ \beta_p \end{bmatrix} = x_i^{\tau} \cdot \beta,$$

$\beta_1, \ldots, \beta_p$ δεν είναι γνωστά και πρέπει να εκτιμηθούν.

$$X = \begin{bmatrix} x_1^{\tau} \\ \cdots \\ x_N^{\tau} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{bmatrix}, \quad x_i = \begin{bmatrix} x_{i1} \\ \cdots \\ x_{ip} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \cdots \\ \beta_p \end{bmatrix}, \quad i = 1, \ldots, N$$

Λειτουργία συνδέσμου

Η συνάρτηση σύνδεσης είναι η g τέτοια ώστε $g(\mu_i) = x_i^{\tau} \cdot \beta$ όπου $\mu_i = E(Y_i)$, $x_i = \begin{bmatrix} x_{i1} \\ \cdots \\ x_{ip} \end{bmatrix}$, $i = 1, \ldots,$ N . Η συνάρτηση g θεωρείται ότι είναι γνωστή και την επιλέγει ο ερευνητής.

Τυχαίο συστατικό

Το τυχαίο συστατικό είναι η κατανομή που επιλέγεται, την οποία ακολουθούν οι μεταβλητές απόκρισης Y_i. Η κατανομή πρέπει να ανήκει στην εκθετική οικογένεια κατανομών. Έστω Y τυχαία ανεξάρτητη μεταβλητή που εξαρτάται από μια παράμετρο θ. Μια κατανομή πιθανότητας ανήκει στην εκθετική οικογένεια κατανομών αν μπορεί να γραφτεί στη μορφή

$f(y; \theta) = \exp(\alpha(y) \cdot s(\theta) + c(\theta) + d(y))$,

όπου α, s, c, d είναι γνωστές συναρτήσεις, ενώ το θ είναι άγνωστο.
Αν $\alpha(y) = y$ η κατανομή λέγεται ότι έχει κανονική μορφή, ενώ s(θ) μερικές φορές λέγεται φυσική παράμετρος της κατανομής.
Αν $g(\mu_i) = \mu_i$ τότε η συνάρτηση σύνδεσης είναι η ταυτοτική, ενώ αν $g(\mu_i) = s(\theta)$ τότε η συνάρτηση σύνδεσης λέγεται κανονική σύνδεση.

Πίνακας 11: Εξίσωση εκθετικής οικογένειας κατανομών

Εκθετική Οικογένεια Κατανομών	
$f(y; \theta) = \exp(\alpha(y) \cdot s(\theta) + c(\theta) + d(y))$	
$\alpha(y) = y$	Κανονική μορφή
$g(\mu_i) = \mu_i$	Συνάρτηση σύνδεσης - ταυτοτική
$g(\mu_i) = s(\theta)$	Συνάρτηση σύνδεσης -κανονική σύνδεση

Πολλές γνωστές κατανομές ανήκουν στην εκθετική οικογένεια. Για παράδειγμα η Poisson, η κανονική, η αρνητική διωνυμική κατανομή μπορούν όλες να γραφούν στην κανονική μορφή $f(y; \theta) = \exp(y \cdot s(\theta) + c(\theta) + d(y))$.

Παράδειγμα εκθετικής οικογένειας κατανομών Poisson και κανονική.

Η πιο απλή κατανομή που χρησιμοποιείται για μοντελοποίηση δεδομένων καταμέτρησης είναι η κατανομή Poisson με συνάρτηση πυκνότητας πιθανότητας για διακριτή τυχαία μεταβλητή $f(y, \mu) = \frac{\mu^y \cdot e^{-\mu}}{y!}$, η οποία έχει τύπο της μορφής $f(y; \theta) = \exp(y \cdot s(\theta) + c(\theta) + d(y))$ και άρα η Poisson παλινδρόμηση είναι μια ειδική περίπτωση του γενικευμένου γραμμικού μοντέλου. Η συνάρτηση $f(y, \mu) = \frac{\mu^y \cdot e^{-\mu}}{y!}$ γράφεται $f(y;\mu) = \exp(y \cdot \log\mu - \mu - \log y!)$ που είναι κανονική μορφή με φυσική παράμετρο $s(\theta) = \log\mu$. Επίσης $c(\theta) = -\mu$, $d(y) = -\log y!$, $\theta = \mu$. Η φυσική παράμετρος για αυτή την οικογένεια είναι $s(\mu) = \log\mu$. Στην γραμμική παλινδρόμηση Poisson η συνάρτηση σύνδεσης είναι η κανονική σύνδεση άρα η συνάρτηση συνδεσης είναι η $g(\mu) = \log\mu$, και το μοντέλο είναι $\log\mu_i = \sum_{j=1}^{p} \beta_j \cdot x_{ij}$, i= 1,…., N

Πίνακας 12: Στην Poisson η συνάρτηση σύνδεσης είναι η log

$s(\theta) = \log\mu$	Συνάρτηση σύνδεσης Poisson, κανονική σύνδεση

Πίνακας 13: Poisson, κανονική, διωνυμική είναι κατανομές της εκθετικής οικογένειας κατανομών.

Κατανομή	Φυσική παράμετρος s	c	d
Poisson	$\log\theta$	$-\theta$	$-\log y!$
Κανονική	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$	$-\frac{y^2}{2\sigma^2}$
Διωνυμική	$\log(\frac{\pi}{1-\pi})$	$n\log(1-\pi)$	$\log\binom{n}{y}$

Στην κανονική κατανομή η συνάρτηση πυκνότητας πιθανότητας είναι $f(y;\ \mu) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \cdot \exp(\ -\ \frac{1}{2\sigma^2} \cdot (y-\mu)^2\ )$, όπου μ είναι η παράμετρος ενδιαφέροντος και σ^2 θεωρείται ως παράμετρος ενόχληση. Αυτό γράφεται $f(y;\ \theta) = \exp(-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\cdot\log(2\pi\sigma^2))$. Αυτή είναι η κανονική μορφή. Η φυσική παράμετρος είναι s $(\mu) = \frac{\mu^2}{\sigma^2}$ και οι άλλοι όροι στην εξίσωση $f(y;\ \theta) = \exp(y\cdot s(\theta)+c(\theta) +d(y))$ είναι $c(\mu) = -\ \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\cdot\log(2\pi\sigma^2)$ και $d(y) = -\ \frac{y^2}{2\sigma^2}$ (Dobson, 2002).

4.4 Επέκταση του γενικευμένου γραμμικού μοντέλου

Η κατανομή σε ένα γενικευμένο γραμμικό μοντέλο έχει εξίσωση της μορφής $f(y_i;\ \theta_i) = \exp(y_i\cdot s(\theta_i) + c(\theta_i) + d(y_i))$. Σε ορισμένες περιπτώσεις και για μεγαλύτερη ακρίβεια είναι χρήσιμο να μπει μια πρόσθετη παράμετρος που μετρά την διασπορά. Για παράδειγμα στην Poisson πρέπει η διακύμανση να είναι ίση με το μέσο, ενώ αυτό μπορεί να μην ισχύει. Με την νέα παράμετρο διασποράς φ η συνάρτηση πυκνότητας πιθανότητας ή μάζας γράφεται

$f(y_i, \theta_i, \varphi) = \exp\left(\frac{y_i \cdot \theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i,\varphi)\right)$,

όπου θ είναι κανονική παράμετρος και φ είναι η παράμετρος διασποράς η οποία δίνεται. Οι συναρτήσεις b( ) και c( ) είναι γνωστές και καθορίζουν ποια κατανομή χρησιμοποιείται, δηλαδή κανονική ή διωνυμική ή Poisson κατανομή. Ο μέσος και η διακύμανση του y_i θεωρούνται τα $E(y_i) = \mu_i = b'(\theta_i)$ και $Var(y_i) = \varphi \cdot b''(\theta_i)$. Όταν φ είναι γνωστό τότε η συνάρτηση είναι ίδια με τον αρχικό τύπο $f(y; \theta) = \exp(y \cdot s(\theta) + c(\theta) + d(y))$ και η μέθοδος είναι ίδια. Συνήθως το φ είναι άγνωστο, και σε αυτές τις περιπτώσεις πρώτα εκτιμάται το φ και μετά ακολουθεί η γνωστή λειτουργία του γενικευμένου γραμμικού μοντέλου (Νταϊλιάνας, 2012).

4.5 Γενικευμένο γραμμικό μοντέλο Poisson για μοντελοποίηση στο ποδόσφαιρο

Η καταμέτρηση δεδομένων όπως ο αριθμός των γκολ για την εντός έδρας και εκτός έδρας ομάδα μπορεί να μοντελοποιηθεί χρησιμοποιώντας Poisson παλινδρόμηση, όπου η μεταβλητή απόκρισης είναι Poisson κατανομή, διότι ο αριθμός των γκολ εντός και εκτός μπορεί να θεωρηθεί περίπου Poisson. Στην περίπτωση Poisson η συνάρτηση σύνδεσης πρέπει πρώτα να οριστεί ξαναγράφοντας τη συνάρτηση μάζας πιθανότητας ως οικογένεια εκθετικής διασποράς. Ο γενικός τύπος μιας εκθετικής οικογένειας διασποράς με N παρατηρήσεις $y_1, .., y_N$ της μεταβλητής απόκρισης Y_i, i=1,…, N με συνάρτηση μάζας πιθανότητας ή πυκνότητας $f(y_i, \theta_i, \varphi)$ για y_i είναι $f(y_i, \theta_i, \varphi) = \exp\left(\frac{y_i \cdot \theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i,\varphi)\right)$. Στην οικογένεια εκθετικής διασποράς θ_i είναι κανονική παράμετρος που εξαρτάται από ένα μοντέλο γραμμικής πρόβλεψης. Η παράμετρος διασποράς είναι φ και δίνεται. Χρησιμοποιώντας την συνάρτηση μάζας πιθανότητας της Poisson κανανομής, ο τύπος της εκθετικής διασποράς είναι:

$$f(y_i, \lambda_i) = \frac{\lambda_i^{y_i} \cdot e^{-\lambda_i}}{y_i!} = \exp\left(\log(e^{-\lambda_i}) + \log(\lambda_i)^{y_i} - \log(y_i!)\right) =$$

$$= \exp\left(-\lambda_i + y_i \cdot \log\lambda_i - \log(y_i!)\right) = \exp\left(y_i \cdot \log\lambda_i - \lambda_i - \log(y_i!)\right)$$

Άρα

$\alpha(\varphi)=1$, $\theta_i = \log(\lambda_i)$, $b(\theta_i) = \lambda_i$, $c(y_i,\varphi) = -\log(y_i!)$

άρα $b(\theta_i) = \lambda_i = e^{\theta i}$

Η αναμενόμενη τιμή και η διακύμανση του Y_i είναι

$E(Y_i) = \mu_i = b'(\theta_i)$ και $Var(Y_i) = b''(\theta_i) \cdot \alpha(\varphi)$

Αφού $b(\theta_i) = \lambda_i$ και $\alpha(\varphi)=1$ στην περίπτωση της Poisson η αναμενόμενη τιμή και η διακύμανση είναι $E(Y_i) = b'(\theta_i) = e^{\theta_i} = \lambda_i$ και $Var(Y_i) = b''(\theta_i) \cdot \alpha(\varphi) = e^{\theta_i} = \lambda_i$.

Άρα η αναμενόμενη τιμή είναι ίση με την διακύμανση όπως αναμενόταν στην Poisson περίπτωση. Η συνάρτηση σύνδεσης είναι η log με $g(\lambda_i) = \theta_i = \log(\lambda_i)$. Άρα ο log της αναμενόμενης τιμής δηλαδή $\log(\lambda_i)$ μπορεί να μοντελοποιηθεί από τους παράγοντες πρόβλεψης x_{ij} όπου β_j είναι ο εκτιμώμενος συντελεστής για κάθε παράγοντα πρόβλεψης. Αυτό εκφράζεται ως $\log(\lambda_i) = \sum_{j=1}^{p} \beta_j . x_{ij}$, i = 1,….,N

Η ένταση του σκοραρίσματος για την εντός έδρας ομάδα λ_x και την εκτός έδρας ομάδα λ_y μπορούν να μοντελοποιηθούν από την Poisson παλινδρόμηση θεωρώντας ως παράγοντες πρόβλεψης την δύναμη επίθεση_x, άμυνα_y και επίθεση_y, άμυνα_x αντίστοιχα. Χρησιμοποιώντας το σύνδεσμο log του γενικευμένου γραμμικού μοντέλου, η ένταση του σκορ για την ομάδα εντός i εναντίον της j ομάδας εκτός μοντελοποιείται ως

$$\log(\lambda_{x,i}) = \beta_0 + \beta_1 \cdot \text{επίθεση}_{x,i} + \beta_2 \cdot \text{άμυνα}_{y,j}$$
$$\log(\lambda_{y,j}) = \beta_0' + \beta_1' \cdot \text{επίθεση}_{y,j} + \beta_2' \cdot \text{άμυνα}_{x,i}$$

4.6 Γενικευμένο γραμμικό μοντέλο αρνητικής διωνυμικής για μοντελοποίηση στο ποδόσφαιρο

Όπως στην περίπτωση της Poisson το γενικευμένο γραμμικό μοντέλο με αρνητική διωνυμική μπορεί να προσαρμοστεί για καταμέτρηση δεδομένων όπως ο αριθμός των γκολ για την ομάδα εντός και εκτός έδρας. Μια συνάρτηση σύνδεσης μπορεί επίσης να βρεθεί για την αρνητική διωνυμική κατανομή. Η συνάρτηση σύνδεσης στην αρνητική διωνυμική περίπτωση αποδεικνύεται ότι είναι $g(\mu_i) = \theta_i = \log \frac{\mu_i}{\mu_i + k} = x_i \cdot \beta$. Επειδή $\mu_i > 0$ προκύπτει ότι $g(\mu_i) < 0$. Ο κανονικός σύνδεσμος δεν ταιριάζει όσον αφορά την μοντελοποίηση των γκολ διότι το μοντέλο πρέπει να μπορεί να παίρνει θετικές τιμές. Μια επιλογή είναι να χρησιμοποιηθεί log σύνδεσμος παρόμοιος με αυτόν στο μοντέλο Poisson. Σε αυτή την περίπτωση το log της αναμενόμενης τιμής μπορεί να μοντελοποιηθεί με τον ίδιο τρόπο όπως στην Poisson (Liden, 2016).

Δηλαδή

$$\log(\mu_{x,i}) = \beta_0 + \beta_1 \cdot \text{επίθεση}_{x,i} + \beta_2 \cdot \text{άμυνα}_{y,j}$$
$$\log(\mu_{y,j}) = \beta_0' + \beta_1' \cdot \text{επίθεση}_{y,j} + \beta_2' \cdot \text{άμυνα}_{x,i}$$

4.7 Η επιλογή της κατάλληλης κατανομής

Συνήθως εξετάζεται το είδος των δεδομένων προκειμένου να καθοριστεί ποια κατανομή θα επιλεγεί ως κατάλληλη για ένα μοντέλο.

Πίνακας 14: Πότε χρησιμοποιείται κάθε κατανομή

Κατανομή	Πότε χρησιμοποιείται
Poisson	Για καταμέτρηση δεδομένων και η κατανομή είναι λοξή διακριτή. Διακριτό αποτέλεσμα σημαίνει ότι οι τιμές είναι ακέραιοι
Logistic	Όταν είναι δυαδικό αποτέλεσμα-δύο επίπεδα μεταβλητής που πρέπει να προβλεφθούν. Παντρεμένος – όχι παντρεμένος, πέρασες τις εξετάσεις-δεν πέρασες
Αρνητική διωνυμική	Όμοια με Poisson χωρίς να απαιτείται η υπόθεση διακύμανση ίση με μέσο

Ο αριθμός των φορών που ένα γεγονός συμβαίνει είναι ένας συνηθισμένος τύπος δεδομένων. Η καταμέτρηση μπορεί να είναι συχνότητες δεδομένων που εμφανίζονται σε ένα πίνακα έκτακτης ανάγκης ή αριθμός γεγονότων όπως τα γκολ τα οποία πρέπει να αναλυθούν σε σχέση με μερικές μεταβλητές όπως η δύναμη της ομάδας ή η έδρα της ομάδας. Για μοντελοποίηση δεδομένων καταμέτρησης συχνά χρησιμοποιείται η Poisson κατανομή και λογαριθμικά γραμμικά μοντέλα.

Το συνηθισμένο γραμμικό μοντέλο είναι μια ειδική περίπτωση του γενικευμένου γραμμικού μοντέλου. Το γραμμικό μοντέλο έχει: συστηματικό συστατικό το $\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots$, συνάρτηση σύνδεσης την ταυτοτική και κατανομή την κανονική.

Η εφαρμογή του γενικευμένου γραμμικού μοντέλου γίνεται εύκολα με την βοήθεια του υπολογιστή, με χρήση του στατιστικού πακέτου R ή SPSS.

Πίνακας 15: Συναρτήσεις οι οποίες έχουν η κάθε μια και άλλο προεπιλεγμένο κώδικα

Όνομα	Συνάρτηση Σύνδεσης	Κατανομή

Γραμμικό μοντέλο	Ταυτοτική $\beta_0 + \beta_1 \cdot x$	Κανονική
Λογιστική παλινδρόμηση	logit $\frac{e^{\beta_0 + \beta_1 \cdot x}}{1+e^{\beta_0 + \beta_1 \cdot x}}$	Διωνυμική
Poisson	log $e^{\beta_0 + \beta_1 \cdot x}$	Poisson

4.8 Γραμμικό μοντέλο

Η πιο απλή περίπτωση του γενικευμένου γραμμικού μοντέλου είναι το γραμμικό μοντέλο $E(Y_i) = \mu_i = x_i^\tau \cdot \beta$, $Y_i \sim N(\mu, \sigma^2)$ όπου $Y_1, \ldots, Y_N$ είναι ανεξάρτητα. Εδώ η συνάρτηση σύνδεσης είναι η ταυτοτική συνάρτηση $g(\mu_i) = \mu_i$. Αυτό το μοντέλο συνήθως γράφεται στη μορφή $y = X \cdot \beta + e$ όπου

$$y = \begin{bmatrix} Y_1 \\ \ldots \\ Y_N \end{bmatrix}, \quad X = \begin{bmatrix} x_1^\tau \\ \ldots \\ x_N^\tau \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \ldots \\ \beta_p \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ \ldots \\ e_N \end{bmatrix}$$

όπου e_i είναι ανεξάρτητα, πανομοιότυπα ανεξάρτητες τυχαίες μεταβλητές με $e_i \sim N(0, \sigma^2)$ για $i = 1, 2, \ldots, N$. Πολλαπλή γραμμική παλινδρόμηση, ανάλυση διακύμανσης anova και ανάλυση συνδιακύμανσης είναι όλα αυτού του τύπου και μαζί μερικές φορές λέγονται γενικά γραμμικά μοντέλα (Dobson, 2002).

Κεφάλαιο 5 Πρόβλεψη αποτελεσμάτων αγώνων ποδοσφαίρου

5.1 Μοντελοποίηση της δύναμης της ομάδας

Ίσως το πιο δύσκολο κομμάτι στην εκτίμηση του σκορ είναι η μοντελοποίηση της ικανότητας των ομάδων. Αυτό κυρίως διότι βάζοντας ένα αριθμό στις ικανότητες είναι αυθαίρετο. Οι διαφορετικές ικανότητες μιας ομάδας καθορίζονται από πολλούς παράγοντες και παραμέτρους και για αυτό τον λόγο έχουν αναπτυχθεί περίπλοκα μοντέλα. Ωστόσο είναι δύσκολο να μετρηθούν όλοι αυτοί οι παράγοντες. Το ζητούμενο είναι το μοντέλο να ενσωματώνει τα πιο σημαντικά χαρακτηριστικά, και το πιο σημαντικό στατιστικό στο ποδόσφαιρο είναι τα γκολ. Τα γκολ σε ένα αγώνα κρίνουν το αποτέλεσμα και είναι ο πιο σημαντικός παράγοντας. Το πιο απλό μοντέλο που εκτιμάει τα γκολ σε ένα αγώνα είναι το μοντέλο του Maher (1982) και το μόνο που χρειάζεται είναι τις ομάδες και τα γκολ παλαιότερων αγώνων. Παρά το γεγονός ότι είναι απλό, η εκτίμηση που κάνει είναι πολύ καλή.

Σύμφωνα με τον Maher η ικανότητα μιας ομάδας περιγράφεται καλύτερα αν θεωρηθούν διαφορετικά μέτρα για την επιθετική και την αμυντική της ικανότητα. Επίσης η επίδραση της έδρας είναι σημαντικός παράγοντας, για αυτό η επιθετική ικανότητα μιας ομάδας θεωρείται ότι είναι διαφορετική όταν παίζει εντός και όταν παίζει εκτός έδρας. Όμοια η αμυντική ικανότητα μιας ομάδας είναι διαφορετική για τα εντός και εκτός έδρας παιχνίδια. Επίσης οι δείκτες της αμυντικής και επιθετικής ικανότητας της ομάδας προκύπτουν από πρόσφατα αποτελέσματα.

Σύμφωνα με τον Maher (1982) οι παράμετροι για την ικανότητα μιας ομάδας είναι οι εξής:

- Δύναμη επίθεσης όταν η ομάδα παίζει εντός έδρας.
- Δύναμη επίθεσης όταν η ομάδα παίζει εκτός έδρας.
- Δύναμη άμυνας όταν η ομάδα παίζει εντός έδρας.
- Δύναμη άμυνας όταν η ομάδα παίζει εκτός έδρας.

Ο αναμενόμενος αριθμός των γκολ της εντός έδρας ομάδας θα εξαρτάται από την δύναμη της επίθεσης της εντός έδρας ομάδας και την δύναμη της άμυνας της εκτός έδρας ομάδας. Αντίστοιχα ο αναμενόμενος αριθμός γκολ της ομάδας που παίζει εκτός έδρας θα εξαρτάται από την δύναμη της επίθεσης της εκτός έδρας ομάδας και την δύναμη της άμυνας της εντός έδρας ομάδας.

Παράδειγμα για τη δύναμη

Εφαρμόζοντας γραμμική παλινδόμηση θα επιχειρηθεί να αποδειχθεί ότι ο αριθμός των γκολ είναι το πιο σημαντικό χαρακτηριστικό της δύναμης της ομάδας. Δηλαδή οι δείκτες

απόδοσης που εξαρτώνται μόνο από τα γκολ, και όχι από άλλους παράγοντες, είναι αντιπροσωπευτικοί. Προκειμένου να υποστηριχθεί αυτή τη δήλωση, υπολογίστηκε η στατιστική σχέση μεταξύ της τελικής βαθμολογίας και του αριθμού των γκολ που βάζει και δέχεται κάθε ομάδα, από μια σεζόν του Ελληνικού πρωταθλήματος.

Στο Ελληνικό πρωτάθλημα ποδοσφαίρου ανταγωνίζονται δεκατέσσερις ομάδες, παίζοντας μεταξύ τους δύο φορές, για συνολικά 26 αγώνες. Η ομάδα με τη μεγαλύτερη διαφορά τερμάτων, υπολογιζόμενη ως τον αριθμό των γκολ που σημείωσε μείον τον αριθμό των γκολ που δέχτηκε, θα κερδίσει συχνά περισσότερα παιχνίδια. Η νίκη περισσότερων παιχνιδιών θα οδηγήσει σε υψηλότερους πόντους που κερδίζονται επειδή ένα παιχνίδι νίκης αξίζει 3 πόντους, ένα παιχνίδι ισοπαλίας αξίζει 1 πόντο και ένα παιχνίδι που χάνεται αξίζει 0 πόντους. Άρα θα πρέπει να υπάρχει μια γραμμική σχέση μεταξύ της διαφοράς τερμάτων και των πόντων που κερδίζει μια ομάδα στο τέλος της σεζόν. Για να διερευνηθεί η σχέση εφαρμόζεται απλή γραμμική παλινδρόμηση στο excel με εξαρτημένη μεταβλητή την βαθμολογία και ανεξάρτητη μεταβλητή την διαφορά τερμάτων. Ο πίνακας περιέχει δεδομένα από τη σεζόν 2020-2021 του ελληνικού πρωταθλήματος, την βαθμολογία των ομάδων και τα γκολ που σημείωσαν.

Πίνακας 16: Δεδομένα από τη σεζόν 2020-21 του Ελληνικού πρωταθλήματος

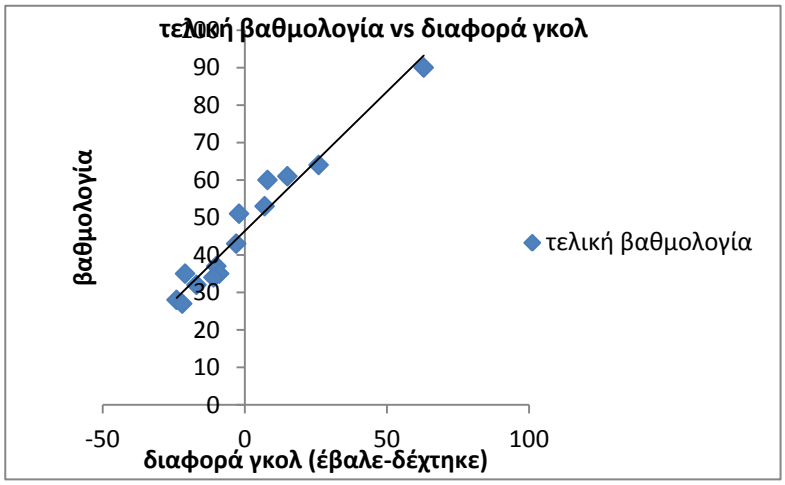
	Βαθμολογία	Γκολ έβαλαν	Γκολ δέχτηκαν	Διαφορά γκολ
Ολυμπιακός	90	82	19	63
ΠΑΟΚ	64	60	34	26
ΑΡΗΣ	61	41	26	15
ΑΕΚ	60	53	45	8
ΠΑΟ	53	41	34	7
Αστέρας	51	36	38	-2
Βόλος	43	34	37	-3
Ατρόμητος	37	30	40	-10

Π.Γιάννενα	35	27	36	-9
Λαμία	35	21	42	-21
Απόλλων	34	29	40	-11
ΟΦΗ	32	30	47	-17
Παναιτωλικός	28	20	44	-24
ΑΕΛ	27	25	47	-22

Από το διάγραμμα διασποράς των δεδομένων φαίνεται ότι τα δεδομένα έχουν γραμμική σχέση και είναι κοντά σε μια ευθεία. Η ομάδα με μικρότερη διαφορά τερμάτων θα έχει μικρότερη βαθμολογία, ενώ αν έχει μεγάλη διαφορά τερμάτων μεγαλύτερη βαθμολογία. Το μοντέλο είναι το εξής:

Συνολική βαθμολογία = β_0+ β_1 ·Διαφορά γκολ + ε

Υπολογίζοντας τους συντελεστές β_0, β_1 μπορεί να εκτιμηθεί η τελική βαθμολογία μιας ομάδας από τα γκολ που έβαλε και δέχτηκε.



Γράφημα 9: Γραμμική παλινδρόμηση, βαθμολογία σε συνάρτηση με τη διαφορά των γκολ

Από το excel και ακολουθώντας τα βήματα data analysis- regression προκύπτουν τα παρακάτω αποτελέσματα.

Πίνακας 17: excel γραμμικής παλινδρόμησης

	Coefficients
Intercept	46,42857143
διαφορα γκολ	0,74280474

$Y = 46{,}42 + 0{,}74 \cdot X$

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0,975542
R Square	0,951681
Adjusted R Square	0,947655
Standard Error	4,067776
Observations	14

Ο παραπάνω πίνακας δείχνει ότι το 97,55% της μεταβλητότητας της βαθμολογίας μιας ομάδας εξηγείται από την μεταβλητότητα της διαφοράς των γκολ που πετυχαίνει. Άρα η συσχέτιση των μεταβλητών είναι υψηλή και ως εκ τούτου οι δείκτες απόδοσης με κριτήριο μόνο τα γκολ είναι ικανοί να εκφράσουν την απόδοση της ομάδας. Τα αποτελέσματα αυτά αν και βασίζονται μόνο σε μια σεζόν του ελληνικού πρωταθλήματος, είναι ενδεικτικά για το ότι ο αριθμός των γκολ που βάζει μια ομάδα είναι η πιο σημαντική παράμετρος της δύναμής της.

5.2 Τα τρία διαφορετικά μοντέλα

Χρησιμοποιώντας παλαιότερα δεδομένα από το Ελληνικό πρωτάθλημα θα συγκριθούν και θα αξιολογηθούν τρία διαφορετικά μοντέλα που προβλέπουν αγώνες ποδοσφαίρου. Τα μοντέλα είναι τα εξής: Το απλό Poisson μοντέλο, το γενικευμένο γραμμικό μοντέλο με κατανομή Poisson, το γενικευμένο γραμμικό μοντέλο με αρνητική διωνυμική. Οι παράγοντες

πρόβλεψης επίθεση$_x$, άμυνα$_y$, επίθεση$_y$, άμυνα$_x$, υπολογίζονται με τον ίδιο τρόπο και στα τρία μοντέλα. Το πρώτο μοντέλο, το απλό Poisson μοντέλο, χρησιμοποιείται σε ιστοσελίδες πονταρίσματος για το ποδόσφαιρο και η εκτίμηση των αναμενόμενων γκολ λ_x και λ_y γίνεται απευθείας από ένα τύπο ή ως ο μέσος όρος του αριθμού των γκολ. Αντίθετα στα γενικευμένα γραμμικά μοντέλα τα λ_x και λ_y υπολογίζονται με παλινδρόμηση, με την βοήθεια του λογισμικού SPSS.

Αφού εκτιμηθούν τα λ_x και λ_y, υπολογίζεται η πιθανότητα του τελικού αποτελέσματος. Αυτό θα γίνει από την συνάρτηση διμεταβλητής πυκνότητας $P(X=x,Y=y) = P(X=x)\cdot P(Y=y)$ για ανεξάρτητες μεταβλητές. Πιο σύνθετα μοντέλα τα οποία επιτρέπουν εξάρτηση μεταξύ των μεταβλητών χρησιμοποιούν άλλους τύπους υπολογισμού των πιθανοτήτων.

Θα ελεγχθεί το απλό Poisson μοντέλο για το αν προσαρμόζεται καλά στα δεδομένα καθώς και για την ανεξαρτησία των γκολ. Τέλος θα προσαρμοστούν τα δύο γενικευμένα γραμμικά μοντέλα και θα αξιολογηθούν με σκοπό να βρεθεί το καλύτερο μοντέλο για τα γκολ.

5.3 Δεδομένα από το Ελληνικό πρωτάθλημα

Αυτό που ενδιαφέρει είναι πόσα γκολ μπήκαν κατά μέσο όρο από τις ομάδες για τα παιχνίδια εντός έδρας και για τα παιχνίδια εκτός έδρας. Ο μέσος αριθμός γκολ που μπήκαν στην έδρα των ομάδων είναι το άθροισμα όλων των γκολ εντός έδρας δια το πλήθος των αγώνων. Όμοια για το μέσο αριθμό γκολ στα εκτός έδρας παιχνίδια. Ο πίνακας δείχνει τους μέσους και τις διακυμάνσεις για τις σεζόν 2016-2017 έως 2019-2020 του Ελληνικού πρωταθλήματος.

Πίνακας 18: Μέσοι και τις διακυμάνσεις για τις σεζόν 2016-2017 έως 2019-2020 του Ελληνικού πρωταθλήματος

Σεζόν	Μέσος γκολ εντός	Διακύμναση γκολ εντός		Μέσος γκολ εκτός	Διακύμανση γκολ εκτός
2016-17	1,44	1,66		0,88	1,07
2017-18	1,31	1,5		0,89	1
2018-19	1,4	1,7		0,92	1,02
2019-20	1,43	1,56		0,94	0,97

Από τον πίνακα φαίνεται ότι οι μέσοι και οι διακυμάνσεις έχουν τιμές αρκετά κοντά μεταξύ τους. Για αυτό το λόγο είναι αληθοφανές να χρησιμοποιηθεί το μοντέλο με κατανομή Poisson. Ωστόσο η διακύμανση είναι λίγο μεγαλύτερη από το μέσο το οποίο δηλώνει υπερδιασπορά στην περίπτωση της Poisson.

5.4 Το απλό ανεξάρτητο μοντέλο Poisson

Έστω ότι τα γκολ εντός και εκτός έδρας είναι δύο τυχαίες ανεξάρτητες μεταβλητές που ακολουθούν την οριακή Poisson κατανομή. Αν X ο αριθμός των γκολ εντός έδρας και Y ο αριθμός των γκολ εκτός έδρας, τότε

$$X \sim \text{Poisson}(\lambda_x) \quad , \quad Y \sim \text{Poisson}(\lambda_y) \quad \text{και}$$

$$P(X=x) = \frac{(\lambda_x)^x \cdot e^{-\lambda_x}}{x!} \quad , \quad P(Y=y) = \frac{(\lambda_y)^y \cdot e^{-\lambda_y}}{y!}$$

Το τελικό αποτέλεσμα του αγώνα θα ακολουθεί τη διμεταβλητή Poisson συνάρτηση πυκνότητας

$$P(X=x,Y=y) = P(X=x) \cdot P(Y=y) = \frac{(\lambda_x)^x \cdot e^{-\lambda_x}}{x!} \cdot \frac{(\lambda_y)^y \cdot e^{-\lambda_y}}{y!}.$$

Στην πιο απλή περίπτωση τα λ_x και λ_y μπορούν να εκτιμηθούν απευθείας από ένα τύπο ή ακόμα και ως ο μέσος αριθμός των γκολ, και όχι με παλινδρόμηση.

Ένα κοινό μοντέλο που υπάρχει για πρόβλεψη αποτελσμάτων σε πολλές ιστοσελίδες πονταρίσματος εκτιμά απευθείας τον αναμενόμενο αριθμό γκολ που θα βάλει η ομάδα που παίζει εντός έδρας λ_x και τον αναμενόμενο αριθμό γκολ για την ομάδα που παίζει εκτός έδρας λ_y. Τα λ_x και λ_y εκτιμώνται από τους παρακάτω τύπους.

λ_x = (συντελεστής επίθεσης x της εντός έδρας ομάδας)·(συντελεστή άμυνας y της εκτός έδρας αντίπαλης ομάδας)·(συνολικός μέσος αριθμός γκολ που μπήκαν στα παιχνίδια εντός έδρας όλη τη σεζόν)

λ_y = (συντελεστής επίθεσης y της εκτός έδρας ομάδας)·(συντελεστή άμυνας x της αντίπαλης εντός έδρας ομάδα)·(συνολικός μέσος αριθμός γκολ που μπήκαν στα παιχνίδια εκτός έδρας όλη τη σεζόν)

όπου,

Συντελεστής επίθεσης x για την ομάδα εντός = επίθεση_x

$$= \frac{\text{μέσος αριθμός γκολ που έβαλε η ομάδας εντός, στα εντός παιχνίδια της}}{\text{μέσος αριθμός όλων των γκολ που έβαλαν οι εντός συνολικά τη σεζόν}}$$

Συντελεστής επίθεσης y για την ομάδα εκτός = επίθεση_y

$$= \frac{\text{μέσος αριθμός γκολ που έβαλε η ομάδα εκτός,όταν έπαιζε εκτός}}{\text{μέσος αριθμός όλων των γκολ που έβαλαν οι εκτός συνολικά τη σεζόν}}$$

Συντελεστής άμυνας x για την ομάδα εντός = άμυνα_x

$$= \frac{\text{μέσος αριθμός γκολ που δέχτηκε η ομάδα εντός,όταν έπαιζε εντός}}{\text{μέσος αριθμός όλων των γκολ που δέχτηκαν οι εντός συνολικά τη σεζόν}}$$

Συντελεστής άμυνας y για την ομάδα εκτός = άμυνα_y

$$= \frac{\text{μέσος αριθμός γκολ που δέχτηκε η ομάδα εκτός,όταν έπαιζε εκτός}}{\text{μέσος αριθμός όλων των γκολ που δέχτηκαν οι εκτός συνολικά τη σεζόν}}$$

(Liden, 2016).

Για παράδειγμα η ομάδα της ΑΕΚ τη σεζόν 2019-2020 έπαιξε συνολικά 34 αγώνες, 17 εντός έδρας και 17 εκτός έδρας, και είχε τα εξής στατιστικά:

- Στους εντός έδρας αγώνες έβαλε συνολικά 35 γκολ, άρα ο μέσος αριθμός γκολ που έβαλε εντός είναι 35/17 = 2,06. Ο συνολικός μέσος αριθμός γκολ που έβαλαν όλες οι ομάδες όταν έπαιζαν εντός είναι 333/230 = 1,45 όπου 333 το σύνολο των γκολ που μπήκαν από τις ομάδες που έπαιζαν εντός και 230 το σύνολο των αγώνων όλη τη σεζόν. Άρα επίθεση_x = 2,06/1,45 = 1,42 είναι ο συντελεστής επίθεσης για την ΑΕΚ στα εντός έδρας παιχνίδια της.
- Στους εντός έδρας αγώνες δέχτηκε συνολικά 16 γκολ, άρα ο μέσος αριθμός γκολ που δέχτηκε εντός είναι 16/17 = 0,94. Ο συνολικός μέσος αριθμός γκολ που δέχτηκαν όλες οι ομάδες όταν έπαιζαν εντός είναι 218/230 = 0,95, όπου 218 το σύνολο των γκολ που δέχτηκαν οι εντός ομάδες ή το σύνολο των εκτός έδρας γκολ τη σεζόν και 230 το πλήθος των αγώνων τη σεζόν. Άρα άμυνα_x= 0,94/0,95 = 0,99 είναι ο συντελεστής άμυνας της ΑΕΚ στα εντός.
- Στους εκτός έδρας αγώνες έβαλε συνολικά 24 γκολ, άρα μέσος αριθμός γκολ που έβαλε εκτός έδρας είναι 24/17 = 1,41. Ο συνολικός μέσος αριθμός γκολ που έβαλαν όλες οι ομάδες όταν έπαιξαν εκτός είναι 218/230 = 0,95 όπου 218 το σύνολο των γκολ που έβαλαν όταν έπαιξαν εκτός και 230 το πλήθος των αγώνων. Τότε επίθεση_y = 1,41/0,95 = 1,49 ο συντελεστής επίθεσης της ΑΕΚ στα εκτός.
- Στους εκτός έδρας αγώνες δέχτηκε συνολικά 13 γκολ, άρα μέσος αριθμός γκολ που δέχτηκε εκτός έδρας είνια 13/17 = 0,76. Ο συνολικός μέσος αριθμός γκολ που δέχτηκαν όλες οι ομάδες όταν έπαιξαν εκτός έδρας είναι 333/230 = 1,45, όπου 333 είναι το άθροισμα όλων των γκολ που δέχτηκαν οι ομάδες στα εκτός και 230 το πλήθος αγώνων. Τότε άμυνα_y = 0,76/1,45 = 0,53 είναι ο συντελεστής άμυνας της ΑΕΚ στα εκτός έδρας.

Αφού ο μέσος αριθμός γκολ που μπήκαν εντός από την ομάδα εντός έδρας όλη τη σεζόν πρέπει να είναι ίσος με τον μέσο αριθμό γκολ που δέχτηκε η ομάδα όταν έπαιζε εκτός τη σεζόν, ο παρανομαστής του συντελεστή επίθεση_x πρέπει να είναι ίσος με τον παρανομαστή του συντελεστή άμυνα_y. Όμοια ο παρανομαστής του συντελεστή άμυνα_x πρέπει να είναι ίσος με τον παρανομαστή του συντελεστή επίθεση_y.

Στην συνέχεια ο αναμενόμενος αριθμός γκολ για την ομάδα εντός, δηλαδή λ_x, δίνεται από τον τύπο:

$E[X] = \lambda_x = (\text{επίθεση}_x) \cdot (\text{άμυνα}_y) \cdot$ (μέσο αριθμό γκολ που μπήκαν στα εντός παιχνίδια συνολικά τη σεζόν) ,

ενώ ο αναμενόμενος αριθμός γκολ για την ομάδα που παίζει εκτός έδρας είναι λ_y:

$E[Y] = \lambda_y = (\text{επίθεση}_y) \cdot (\text{άμυνα}_x) \cdot$ (μέσο αριθμό γκολ που μπήκαν στα εκτός παιχνίδια συνολικά τη σεζόν)

Για παράδειγμα η ΑΕΚ σύμφωνα με τα παραπάνω στατιστικά αναμένεται να πετύχει λ_x γκολ εντός έδρας και λ_y γκολ εκτός έδρας.

$E[X] = \lambda_x = 1{,}42 \cdot (\text{άμυνα}_y \text{ αντίπαλης}) \cdot 1{,}45$

$E[Y] = \lambda_y = 1{,}49 \cdot (\text{άμυνα}_x \text{ αντίπαλης}) \cdot 0{,}95$

Έχοντας τις αναμενόμενες τιμές λ_x και λ_y για τις δύο τυχαίες μεταβλητές, η διμεταβλητή Poisson κατανομή μπορεί να χρησιμοποιηθεί για να υπολογίσει η πιθανότητα για κάθε τελικό αποτέλεσμα χρησιμοποιώντας την συνάρτηση διμεταβλητής πυκνότητας $P(X=x,Y=y) = P(X=x) \cdot P(Y=y)$.

Παράδειγμα για την εκτίμηση του αποτελέσματος σε ένα μελλοντικό παιχνίδι ΑΕΚ-ΠΑΟΚ. Οι παράμετροι έχουν υπολογιστεί, από παλαιότερα δεδομένα, από τη σεζόν 2019-2020 του ελληνικού πρωταθλήματος.

Πίνακας 19: Παράμετροι από τη σεζόν 2019-2020 του ελληνικού πρωταθλήματος

Εντός	Εκτός	επίθεση_x	άμυνα_y	επίθεση_y	άμυνα_x	λ_x	λ_y
ΑΕΚ	ΠΑΟΚ	1,42	0,65	1,68	0,99	1,34	1,58

- Η πιθανότητα η ΑΕΚ , η οποία παίζει εντός έδρας, να βάλει 0 γκολ στο παιχνίδι με τον ΠΑΟΚ είναι

$$P(X=0) = \frac{(\lambda_x)^x \cdot e^{-\lambda_x}}{x!} = \frac{(1,34)^0 \cdot e^{-1,34}}{0!} = 0,2623 = 26,23\%$$

- Η πιθανότητα ο ΠΑΟΚ , ο οποίος παίζει εκτός έδρας, να βάλει 1 γκολ με την ΑΕΚ είναι

$$P(Y=1) = \frac{(\lambda_y)^y \cdot e^{-\lambda_y}}{y!} = \frac{(1,58)^1 \cdot e^{-1,58}}{1!} = 0,3258 = 32,58\%$$

- Η πιθανότητα να έρθει το παιχνίδι ΑΕΚ- ΠΑΟΚ 0-1 είναι

$P(X=0, Y=1) = P(X=0) \cdot P(Y=1) = 26,23\% \cdot 32,58\% = 8,54\%$

.

- Οι πιθανότητες για την εντός νίκη, ισοπαλία ή εκτός νίκη μπορούν να υπολογιστούν αθροίζοντας τις πιθανότητες από τα διαφορετικά αποτελέσματα. Έστω ότι έχουν υπολογιστεί οι πιθανότητες αποτελεσμάτων από 0-0 εώς 5-5, δηλαδή 36 διαφορετικά αποτελέσματα. Η πιθανότητα η ομάδα να βάλει 6 ή περισσότερα γκολ είναι μικρή και θεωρείται αμελητέα. Σε αυτή την περίπτωση η πιθανότητα να κερδίσει η εντός ομάδα είναι:

$P(X>Y) = P(1,0)+P(2,0)+P(3,0)+P(4,0)+P(5,0)+P(2,1)+\ldots..+P(5,3)+P(5,4).$

5.4.1 Έλεγχος για Poisson κατανομή στο απλό μοντέλο Poisson

Το παραπάνω απλό μοντέλο Poisson υποθέτει ότι και οι δύο οριακές κατανομές είναι Poisson. Αυτή η υπόθεση μπορεί να ελεγχθεί. Χρησιμοποιώντας παλαιότερα δεδομένα από τις σεζόν 1994-95 έως 2019-2020 του ελληνικού πρωταθλήματος γίνεται σύγκριση μεταξύ των παρατηρούμενων και των εκτιμώμενων Poisson συχνοτήτων του αριθμού των γκολ που σημειώθηκαν.

Στο γράφημα 10 φαίνεται ότι οι προσδοκώμενες Poisson συχνότητες είναι κοντά στις πραγματικές συχνότητες για τον αριθμό των γκολ που έβαλε η εντός ομάδα. Η αναμενόμενη τιμή της Poisson κατανομής δηλαδή λ θεωρήθηκε σε αυτό το παράδειγμα ως ο συνολικός μέσος αριθμός των γκολ που έβαλε η εντός ομάδα σε όλες τις σεζόν από 1994-95 έως 2019-2020, δηλαδή το λ ισούται με το άθροισμα όλων των γκολ που μπήκαν εντός έδρας σε όλες τις σεζόν προς το πλήθος των παιχνιδιών. Σύμφωνα με τους Καρλή και Ντζούφρα (2003) για πρακτικούς σκοπούς η δύναμη της επίθεσης μπορεί να θεωρηθεί ως ο μέσος αριθμός γκολ που έβαλε η ομάδα, ενώ η δύναμη της άμυνας ως ο μέσος αριθμός γκολ που δέχτηκε.

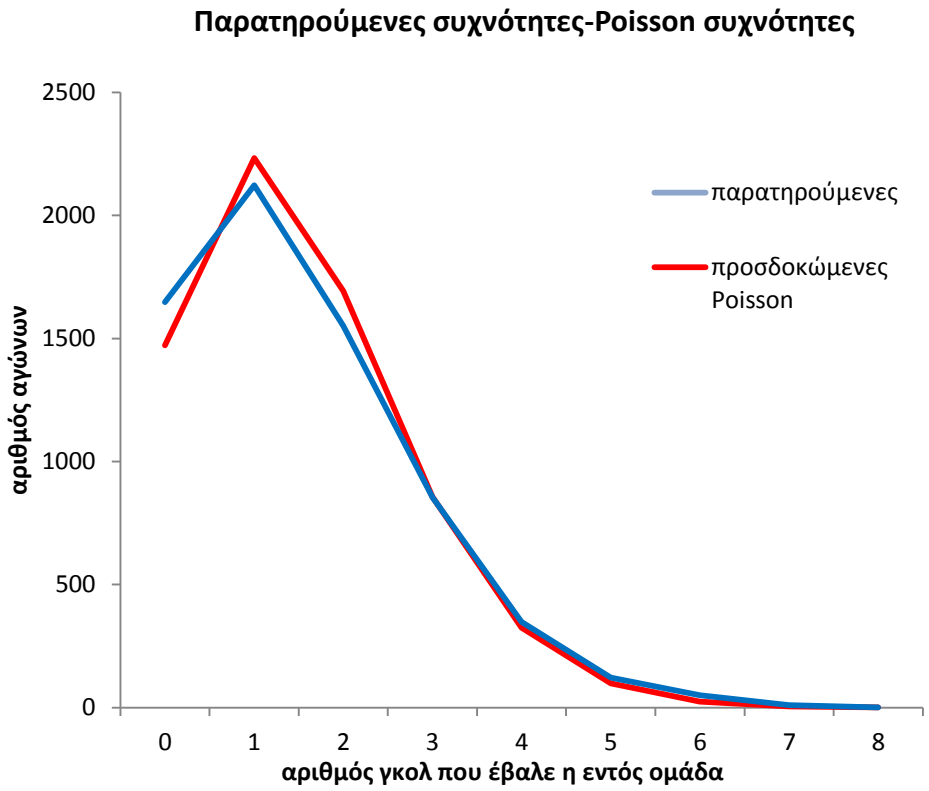
Τα προσδοκώμενα γκολ υπολογίστηκαν ως εξής;

$$\lambda = \frac{\text{άθροισμα όλων των γκολ εντός από 1994 έως 2020}}{\text{πλήθος παιχνιδιών}} = \frac{10172}{6708} = 1,516398$$

$P(X=0) = \frac{(\lambda_x)^x \cdot e^{-\lambda_x}}{x!} = \frac{(1,516398)^0 \cdot e^{-1,516398}}{0!} = 0,219501035.$

Προσδοκώμενα Poisson 0 γκολ= P(X=0)·(πλήθος αγώνων) = 0,2195·6708 =1472,413

Με τον ίδιο τρόπο υπολογίζονται όλα τα προσδοκώμενα γκολ. Όμοια το γράφημα 11 συγκρίνει τις πραγματικές με τις προσδοκώμενες Poisson συχνότητες για τον αριθμό των γκολ που μπήκαν εκτός έδρας.

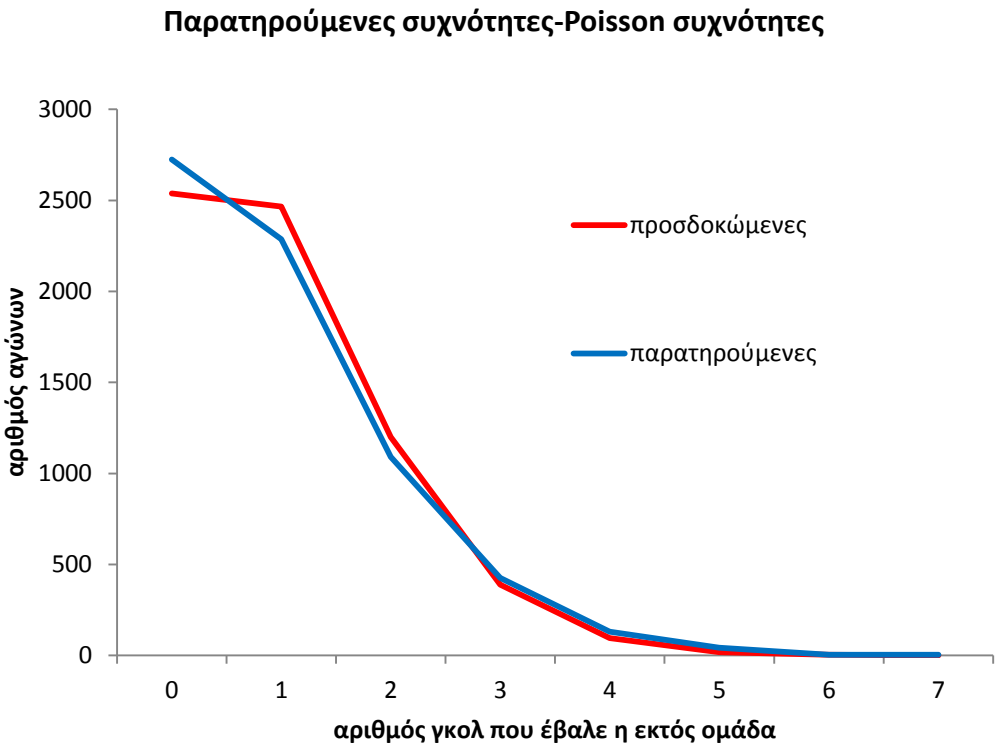


Γράφημα 10: Σύγκριση πραγματικών με προσδοκώμενες Poisson συχνότητες για τον αριθμό των γκολ που μπήκαν στους αγώνες εντός έδρας.

Πίνακας 20: Συγκρίνοντας παρατηρούμενες και αναμενόμενες Poisson συχνότητες για τον αριθμό των γκολ εντός έδρας

Γκολ εντός έδρας	Παρατηρούμενα	Προσδοκώμενα Poisson
0	1648	1472

1	2122	2233
2	1552	1693
3	854	856
4	349	324
5	122	98
6	50	25
7	10	5
8	1	1



Γράφημα 11: Σύγκριση πραγματικών με προσδοκώμενες Poisson συχνότητες για τον αριθμό των γκολ που μπήκαν στους αγώνες εκτός έδρας.

Πίνακας 21: Συγκρίνοντας παρατηρούμενες και αναμενόμενες Poisson συχνότητες για τον αριθμό των γκολ εκτός έδρας.

Γκολ εκτός έδρας	Παρατηρούμενα	Προσδοκώμενα Poisson
0	2725	2538
1	2287	2467
2	1091	1199
3	426	388
4	131	94
5	42	18
6	3	3
7	3	0

Συγκρίνοντας τις παρατηρούμενες και τις προσδοκώμενες Poisson συχνότητες στα γραφήματα και στους πίνακες φαίνεται ότι η Poisson κατανομή προσαρμόζεται αρκετά καλά με τα δεδομένα, αν και η παρατηρούμενη κατανομή έχει μεγαλύτερη ουρά από την Poisson. Το γεγονός αυτό σημαίνει ότι η Poisson παρουσιάζει υπερδιασπορά, δηλαδή η διακύμανση είναι μεγαλύτερη από το μέσο.

5.4.2 Τεστ καλής προσαρμογής Pearson στο απλό μοντέλο Poisson

Θα εξεταστεί κατά πόσο το απλό μοντέλο Poisson προσαρμόζεται καλά στα δεδομένα. Για τον λόγο αυτό θα γίνει ένας χ^2 στατιστικός έλεγχος καλής προσαρμογής στο πρόγραμμα SPSS. Τα αποτελέσματα θα δείξουν αν υπάρχει σημαντική διαφορά μεταξύ των παρατηρούμενων συχνοτήτων Oi και των προσδοκώμενων συχνοτήτων Ei της κατανομής Poisson. Το τεστ υπολογίζει το στατιστικό $X^2 = \sum_{i=1}^{c} \frac{(Oi-Ei)^2}{Ei} \sim x^2_{c-1}$ σε επίπεδο σημαντικότητας 5%, ενώ έχει c-1 βαθμούς ελευθερίας, όπου c είναι ο αριθμός των κελιών. Το τεστ εκτελείται στο SPSS δύο φορές, πρώτα για τα γκολ εντός έδρας και μετά για τα γκολ εκτός έδρας.

Αν p-value είναι μεγαλύτερο από 0,05 τότε δεν είναι στατιστικά σημαντικό σε επίπεδο σημαντικότητας 5% και δεν υπάρχει διαφορά στις παρατηρούμενες και προσδοκώμενες συχνότητες. Αν p-value είναι μικρότερο από 0,05 τότε είναι στατιστικά σημαντικό και σε αυτή την περίπτωση οι παρατηρούμενες συχνότητες είναι διαφορετικές από τις αναμενόμενες συχνότητες. Από το λογισμικό SPSS και από data view τα βήματα είναι τα εξής: Anylyze →Nonparametric Tests→Legacy Dialogs→Chi- square

Chi-Square Test για εντός έδρας γκολ

Πίνακας 22: χ^2 στατιστικός έλεγχος καλής προσαρμογής στο πρόγραμμα SPSS

hgoal			
	Observed N	Expected N	Residual
0	1648	1472,2	175,8
1	2122	2233,3	-111,3
2	1552	1693,3	-141,3
3	854	856,1	-2,1
4	349	324,0	25,0
5	122	98,0	24,0
6	50	25,0	25,0
7	10	5,0	5,0
8	1	1,0	,0
Total	6708		

Test Statistics	
	hgoal
Chi-Square	76,104[a]
df	8
Asymp. Sig.	<,001

Το χ^2 τεστ καταλληλότητας υπολογίζει το άθροισμα των τυποποιημένων καταλοίπων στο τετράγωνο. $X^2 = \sum_ι (\frac{Oι-Eι}{\sqrt{Eι}})^2 = (\frac{1648-1472,2}{\sqrt{1472,2}})^2 + ...+ (\ )^2 \ = 76,104$.

Συγκρίνεται η τιμή $X^2 = 76,104$ με την κρίσιμη τιμή από τον πίνακα της χ^2 κατανομής με df = c-1= 9-1= 8 βαθμούς ελευθερίας, που είναι $\chi^2_{0,05}$ =15,507. Επειδή 76,104 > 15,507, απορρίπτεται η μηδενική υπόθεση.
Εναλλακτικά το συμπέρασμα προκύπτει από το p-value.
Η μηδενική υπόθεση H_0, ότι το Poisson μοντέλο προσαρμόζεται καλά στα δεδομένα, απορρίπτεται διότι p-value < 0,001. Το ίδιο αποτέλεσμα προκύπτει και για τα εκτός έδρας γκολ.

Το συμπέρασμα είναι ότι η υπόθεση Poisson απορρίπτεται τόσο για τα εντός όσο και για τα εκτός έδρας γκολ. Άρα υπάρχει ισχυρή απόδειξη ότι η κατανομή Poisson με το απλό μοντέλο δεν προσαρμόζεται καλά στα δεδομένα. Αυτή η πληροφορία οδηγεί στην αναζήτηση ενός πιο αποτελεσματικού μοντέλου ή μιας άλλης κατανομής ή ενός άλλου υπολογισμού του λ, που να μην επιτρέπει διαφορές μεταξύ πραγματικών και αναμενόμενων συχνοτήτων για τον αριθμό των γκολ.

5.4.3 Έλεγχος ανεξαρτησίας στο απλό μοντέλο Poisson

Η υπόθεση ότι τα γκολ εντός έδρας και εκτός έδρας είναι ανεξάρτητες τυχαίες μεταβλητές ελέγχεται για το αν είναι αληθής, στην περίπτωση του απλού Poisson μοντέλου, με το χ^2 τεστ στατιστικό έλεγχο ανεξαρτησίας του Pearson. Για τον λόγο αυτό χρησιμοποιήθηκαν αποτελέσματα 6709 παιχνιδιών από τις σεζόν 1994 έως 2020 του ελληνικού πρωταθλήματος. Προκύπτει ο παρακάτω πίνακας έκτακτης ανάγκης, ο οποίος απεικονίζει τις συχνότητες παιχνιδιών με ένα συγκεκριμένο αποτέλεσμα. Το τεστ για να είναι αξιόπιστο απαιτεί οι αριθμοί στα κελιά να μην είναι πολύ μικροί, για αυτό τα παιχνίδια με γκολ περισσότερα από 4 έχουν ομαδοποιηθεί ως 4+.

Πίνακας 23: Πίνακας έκτακτης ανάγκης με παρατηρούμενες συχνότητες

Home/Away	0	1	2	3	4+	Total
0	620	529	301	136	62	1648
1	881	705	347	132	57	2122
2	639	561	234	84	34	1552
3	355	302	137	38	22	854
4+	230	190	72	36	5	533
Total	2725	2287	1091	426	180	6709

Χρησιμοποιώντας το λογισμικό R μπορεί να εξεταστεί η μηδενική υπόθεση H_0 ότι οι 2 τυχαίες μεταβλητές, γκολ εντός και γκολ εκτός έδρας, είναι ανεξάρτητες. Το λογισμικό R υπολογίζει τον πίνακα με τις αναμενόμενες συχνότητες καθώς και την τιμή του στατιστικού $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(Oij-Eij)^2}{Eij} \sim \chi^2_{(r-1)(c-1)}$, όπου (r-1)(c-1) είναι οι βαθμοί ελευθερίας.

Εισάγοντας τα δεδομένα στο λογισμικό πρόγραμμα R προκύπτουν τα παρακάτω αποτελέσματα:

```
> chisq.test(t)
```

Pearson's Chi-squared test

data: t

X-squared = 49.468, df = 16, p-value = 2.785e-05

Η παρατηρηθείσα τιμή του στατιστικού ελέγχου είναι $X^2 = 49.468$. Η στατιστική συνάρτηση του ελέγχου κάτω από την μηδενική υπόθεση H_0 ακολουθεί X^2 με 16 βαθμούς ελευθερίας .

Με χρήση της R επίσης

> qchisq(0.95, df = 16)

[1] 26.29623

Επειδή $X^2 = 49.468 > X^2_{0.95,\ 16} = 26.29623$ απορρίπτεται η μηδενική υπόθεση H_o σε επίπεδο στατιστικής σημαντικότητας α = 5%. Ένας άλλος ευκολότερος τρόπος είναι το p–value του ελέγχου. Αν p-value είναι μικρότερο από 0,05 τότε η μηδενική υπόθεση απορρίπτεται. p-value = 2.785e-05 < 0.05, συνεπώς σε επίπεδο στατιστικής σημαντικότητας α = 5% απορρίπτεται η μηδενική υπόθεση H_0 . Το συμπέρασμα είναι ότι δεν υπάρχει στατιστική ανεξαρτησία μεταξύ του αριθμού των εντός και εκτός έδρας γκολ.

Ωστόσο το τεστ μπορεί να δώσει διαφορετικά αποτελέσματα όταν το δείγμα είναι μικρότερο. Η απόρριψη της ανεξαρτησίας μπορεί να μην είναι αξιόπιστη και να οφείλεται στο γεγονός ότι ο πίνακας έκτακτης ανάγκης περιέχει αποτελέσματα από μεγάλο δείγμα παιχνιδιών. Στο συγκεκριμένο παράδειγμα το πλήθος ήταν μεγάλο και ίσο με 6709. Σύμφωνα με τους Καρλή και Ντζούφρα (2003), σε 15 από τα 24 πρωταθλήματα που έγινε έλεγχος ανεξαρτησίας η υπόθεση ανεξαρτησίας δεν απορρίφθηκε σε επίπεδο σημαντικότητας 5%. Αντίθετα η υπόθεση απορρίφθηκε όταν ο πίνακας έκτακτης ανάγκης περιείχε αποτελέσματα όλων των πρωταθλημάτων.

5.5 Γενικευμένο γραμμικό μοντέλο Poisson

Το γενικευμένο γραμμικό μοντέλο Poisson μοντελοποιεί τα αναμενόμενα γκολ εντός έδρας λ_x, χρησιμοποιώντας συνάρτηση σύνδεσης την log. Η μοντελοποίηση έχει τις παραμέτρους $επίθεση_x$ (home attack) και $άμυνα_y$ (away defense) όπως το απλό μοντέλο της προηγούμενης παραγράφου, ως παράγοντες πρόβλεψης. Στη συνέχεια θα γίνει το ίδιο για αναμενόμενα γκολ εκτός έδρας λ_y με παράγοντες πρόβλεψης τις παραμέτρους $επίθεση_y$ (away

attack) και άμυνα$_x$ (home defense). Για όλες τις ομάδες που παίζουν εντός έδρας, προστίθεται ένας ακόμα προβλεπτικός παράγοντας που είναι η επίδρασης έδρας. Το ζητούμενο είναι , με την βοήθεια του λογισμικού SPSS, να βρεθούν οι συντελεστές α, β_0, β_1, β_2 και α΄, β_1΄, β_2΄ στα παρακάτω μοντέλα.

$$\log(\lambda_x) = \alpha + \beta_0 \cdot (\text{επίδραση έδρας}) + \beta_1 \cdot (\text{επίθεση}_x) + \beta_2 \cdot (\text{άμυνα}_y)$$

$$\log(\lambda_y) = \alpha' + \beta_1' \cdot (\text{επίθεση}_y) + \beta_2' \cdot (\text{άμυνα}_x)$$

Ένας συγκεκριμένος αριθμός αγώνων θα χρησιμοποιηθεί. Τα δεδομένα προέρχονται από ένα σύνολο 237 αγώνων της σεζόν 2021-2022 του Ελληνικού πρωταθλήματος. Δεδομένα για την δύναμη επίθεση$_x$, άμυνα$_y$, επίθεση$_y$, άμυνα$_x$, επίδραση έδρας των ομάδων του Ελληνικού πρωταθλήματος τη σεζόν 2021-2022. Η επίδραση της έδρας έχει υπολογιστεί σύμφωνα με την εργασία των Clarke and Norman (1995).

Πίνακας 24: Δεδομένα από τη σεζόν 2021-2022 του Ελληνικού πρωταθλήματος

	Εντός έδρας	Εκτός έδρας	επίθεση$_x$	άμυνα$_y$	επίθεση$_y$	άμυνα$_x$	Επίδραση έδρας
1	Παναιτωλ.	Αστέρ.Τρ.	0,58	0,83	0,7	1,02	0,032052
2	ΠΑΟ	Απόλλ.Σμ	1,54	1,86	0,57	0,41	2,032052
3	Βόλος	Λαμία	1,11	0,79	0,54	1,2	-0,21795
4	ΠΑΟΚ	Γιάννενα	1,3	0,96	0,65	0,97	-0,30128
.....	.....	.....	.....	.....	.....	.....	.....
236	Ιωνικός	Παναιτωλ	1,18	1,38	1,1	1,1	0,365385
237	Λαμία	Ατρόμητ.	0,65	1,4	0,74	1,55	-0,46795

Το λογισμικό πρόγραμμα SPSS δίνει: Το μοντέλο (1)

$$\log(\lambda_x) = \alpha + \beta_0 \cdot (\text{επίδραση έδρας}) + \beta 1 \cdot (\text{επίθεση}_x) + \beta 2 \cdot (\text{άμυνα}_y)$$

Πίνακας 25: Εκτίμηση παραμέτρων γενικευμένο γραμμικό μοντέλο Poisson για τα γκολ εντός έδρας

Model Information

Dependent Variable	homegoal
Probability Distribution	Poisson
Link Function	Log

Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1,903	,2955	-2,482	-1,324	41,473	1	<,001
homeadvantage	-,035	,0811	-,194	,124	,183	1	,669
homeattack	1,168	,2189	,739	1,597	28,479	1	<,001
awaydefense	,910	,1534	,610	1,211	35,226	1	<,001
(Scale)	1[a]						

Dependent Variable: homegoals

Model: (Intercept), homeadvantage, homeattack, awaydefense

a. Fixed at the displayed value.

Από την παραπάνω έξοδο του SPSS το μοντέλο 1 είναι

$$\log(\lambda_x) = -1.903 - 0{,}035 \cdot \text{επίδραση έδρας} + 1{,}168 \cdot \text{επίθεση}_x + 0{,}910 \cdot \text{άμυνα}_y$$

Για παράδειγμα έστω ότι η εντός ομάδα ΑΕΚ πρόκειται να παίξει εναντίον του Ολυμπιακού. Χρησιμοποιώντας το προσαρμοσμένο μοντέλο, ο αναμενόμενος αριθμός γκολ για την ΑΕΚ μπορεί να υπολογιστεί. Σε αυτή την περίπτωση η ΑΕΚ έχει συντελεστή επίθεση_x ίσο με 1,214231, επίδραση έδρας 0,365385 και ο Ολυμπιακός έχει συντελεστή άμυνα_y ίσο με 0,539648. Άρα ο αναμενόμενος αριθμός γκολ για την ΑΕΚ σύμφωνα με το μοντέλο είναι:

$$\log(\lambda_x) = -1.903 - 0{,}035 \cdot \text{home.adv} + 1{,}168 \cdot \text{home.att.} + 0{,}910 \cdot \text{away.def.}$$

$$\log(\lambda_x) = -1.903 - 0{,}035 \cdot 0{,}365385 + 1{,}168 \cdot 1{,}214231. + 0{,}910 \cdot 0{,}539648.$$

$$\log(\lambda_x) = -1.903 - 0{,}0128 + 1{,}4182 + 0{,}4911$$

$$\lambda_x = \exp(-0{,}00652) = 0{,}993501$$

Άρα η εντός έδρας ομάδα ΑΕΚ αναμένεται να σκοράρει 0,993501 γκολ εναντίον της ομάδας του Ολυμπιακού. Τα αναμενόμενα γκολ για την εκτός ομάδα μοντελοποιούνται με τον ίδιο τρόπο αλλά τώρα με τους συντελεστές άμυνα$_x$ της εντός ομάδας και την επίθεση$_y$ της εκτός ομάδας ως προβλεπτικοί παράγοντες. Χρησιμοποιώντας τους ίδιους 237 αγώνες ως δεδομένα πρόβλεψης προσαρμόζεται ένα γενικευμένο γραμμικό μοντέλο για τον αριθμό των γκολ εκτός έδρας που δίνει το παρακάτω μοντέλο.

Προσαρμόζεται το μοντέλο στο SPSS

Το μοντέλο (2) $\log(\lambda_y) = \alpha' + \beta 1' \cdot \text{επίθεση}_y + \beta 2' \cdot \text{άμυνα}_x$

Πίνακας 26: Εκτίμηση παραμέτρων γενικευμένο γραμμικό μοντέλο Poisson για τα γκολ εκτός έδρας

Model Information

Dependent Variable	awaygoals
Probability Distribution	Poisson
Link Function	Log

Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-2,201	,3387	-2,865	-1,537	42,216	1	<,001
awayattack	1,056	,1950	,674	1,438	29,314	1	<,001
homedefense	1,094	,2210	,660	1,527	24,486	1	<,001

(Scale)	1[a]						

Dependent Variable: awaygoals

Model: (Intercept), awayattack, homedefense

a. Fixed at the displayed value.

Από την παραπάνω έξοδο του SPSS το μοντέλο 2 είναι

$$\log(\lambda_y) = -2{,}201 + 1{,}056 \cdot \text{επίθεση}_y + 1{,}094 \cdot \text{άμυνα}_x$$

Για τον αγώνα ΑΕΚ-Ολυμπιακός ο αναμενόμενος αριθμός γκολ εκτός δηλαδή τα αναμενόμενα γκολ του Ολυμπιακού μπορούν να υπολογιστούν από το μοντέλο. Αφού η ΑΕΚ έχει συντελεστή άμυνα_x εντός 0,83647 και ο Ολυμπιακός έχει συντελεστή επίθεση_y 1,5056 ο αναμενόμενος αριθμός γκολ για τον Ολυμπιακό είναι

$$\log(\lambda_y) = -2{,}201 + 1{,}056 \cdot \text{away.attack} + 1{,}094 \cdot \text{home.defence}$$

$$\lambda_y = \exp(-2{,}201 + 1{,}056 \cdot 1{,}5056 + 1{,}094 \cdot 0{,}83647)$$

$$\lambda_y = \exp(0{,}3040117) = 1{,}355285$$

Άρα ο Ολυμπιακός αναμένεται να βάλει 1,355285 γκολ εκτός έδρας με την ΑΕΚ. Ο αναμενόμενος αριθμός γκολ για την εντός ομάδα ΑΕΚ είναι 0,993501. Αφού υπολογίστηκαν τα αναμενόμενα γκολ, από την κοινή συνάρτηση μάζας πιθανότητας, υπολογίζονται οι πιθανότητες νίκη, ισοπαλία ή ήττα.

5.6 Γενικευμένο γραμμικό μοντέλο αρνητικής διωνυμικής

Ένα γενικευμένο γραμμικό μοντέλο αρνητικής διωνυμικής προσαρμόζεται στα ίδια παιχνίδια του ελληνικού πρωταθλήματος για τη σεζόν 2021-2022, και χρησιμοποιεί τα ίδια δεδομένα του πίνακα με τους συντελεστές επίθεση_x (εντός) και άμυνα_y (εκτός) όπως με το γενικευμένο γραμμικό Poisson. Η διαδιακασία είναι ακριβώς η ίδια, το μόνο που αλλάζει είναι η κατανομή. Η έξοδος SPSS δίνει τα αποτελέσματα και τον αναμενόμενο αριθμό γκολ για την εντός ομάδα λ_x και για την εκτός ομάδα λ_y. Τα μοντέλα είναι:

$$\log(\lambda_x) = \gamma + \delta_o \cdot \text{επίδραση έδρας} + \delta_1 \cdot \text{επίθεση}_x + \delta_2 \cdot \text{άμυνα}_y$$

$$\log(\lambda_y) = \gamma' + \delta_1' \cdot \text{επίθεση}_y + \delta_2' \cdot \text{άμυνα}_x$$

Το μοντέλο (3) $\log(\lambda_x) = \gamma + \delta_o \cdot \text{επίδραση έδρας} + \delta_1 \cdot \text{επίθεση}_x + \delta_2 \cdot \text{άμυνα}_y$

Πίνακας 27: Εκτίμηση παραμέτρων γενικευμένο γραμμικό μοντέλο nb για τα γκολ εντός έδρας

Model Information

Dependent Variable	homegoals
Probability Distribution	Negative binomial (1)
Link Function	Log

Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-2,007	,4388	-2,867	-1,146	20,908	1	<,001
homeadvantage	-,055	,1296	-,309	,199	,179	1	,673
homeattack	1,224	,3188	,599	1,848	14,726	1	<,001
awaydefense	,962	,2563	,459	1,464	14,078	1	<,001
(Scale)	1[a]						
(Negative binomial)	1[a]						

Dependent Variable: homegoals

Model: (Intercept), homeadvantage, homeattack, awaydefense

a. Fixed at the displayed value.

Από την παραπάνω έξοδο του SPSS το μοντέλο 3 είναι

$$\log(\lambda_x) = -2{,}007 - 0{,}055 \cdot \text{επίδραση έδρας} + 1{,}224 \cdot \text{επίθεση}_x + 0{,}962 \cdot \text{άμυνα}_y$$

Για τον αγώνα ΑΕΚ- Ολυμπιακός έχουμε

$\log(\lambda_x) = -2{,}007 - 0{,}055 \cdot 0{,}365385 + 1{,}224 \cdot 1{,}214231 + 0{,}962 \cdot 0{,}539648$ $\log(\lambda_x) =$ - 2,007 – 0,0201 + 1,4862 + 0,51914 = - 0,0217586

Άρα λ_x = 0,978476

Το μοντέλο για τα αναμενόμενα γκολ εκτός έδρας χρησιμοποιώντας τους συντελεστές επίθεση_y και άμυνα_x προσαρμόστηκε στο SPSS με αρνητική διωνυμική.

Το μοντέλο (4) $\log(\lambda_y) = \gamma' + \delta_1' \cdot \text{επίθεση}_y + \delta_2' \cdot \text{άμυνα}_x$

Πίνακας 28: Εκτίμηση παραμέτρων γενικευμένο γραμμικό μοντέλο nb για τα γκολ εκτός έδρας

Model Information

Dependent Variable	awaygoals
Probability Distribution	Negative binomial (1)
Link Function	Log

Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-2,293	,4813	-3,237	-1,350	22,702	1	<,001
awayattack	1,079	,2847	,521	1,637	14,370	1	<,001
homedefense	1,159	,3245	,523	1,795	12,746	1	<,001
(Scale)	1[a]						
(Negative binomial)	1[a]						

Dependent Variable: awaygoals

Model: (Intercept), awayattack, homedefense

a. Fixed at the displayed value.

Από το παραπάνω έξοδο του SPSS το μοντέλο 4 είναι

$$\log(\lambda_y) = -2{,}293 + 1{,}079 \cdot \text{επίθεση}_y + 1{,}159 \cdot \text{άμυνα}_x$$

Για τον αγώνα ΑΕΚ – Ολυμπιακός

$\log(\lambda_y) = -2{,}293 + 1{,}079 \cdot 1{,}5056 + 1{,}196 \cdot 0{,}83647 = 0{,}331918$

Άρα $\lambda_y = 1{,}3936$

5.7 Εκτίμηση στοιχηματικών αποδόσεων αγώνα

Στα δεδομένα που χρησιμοποιήθηκαν, για τα γενικευμένα γραμμικά μοντέλα, δεν συμπεριελήφθησαν οι τρεις τελευταίοι αγώνες του πρωταθλήματος. Ο σκοπός είναι ότι θα χρησιμοποιηθούν σε αυτό το παράδειγμα για πρόβλεψη των αποδόσεων των ομάδων.

Θα γίνει πρόβλεψη των αποδόσεων εντός νίκης, ισοπαλίας και εκτός νίκης, όπως υπάρχουν στις εταιρίες στοιχηματισμού, και θα βρεθούν ποιες πρέπει να είναι οι δίκαιες πιθανότητες απόδοσης, σύμφωνα με το γενικευμένο γραμμικό μοντέλο Poisson. Οι πιθανότητες υπολογίζονται θεωρώντας ότι τα γκολ εντός και εκτός είναι ανεξάρτητα. Οι τρεις τελευταίοι αγώνες καθώς και ο αναμενόμενος αριθμός γκολ για την εντός και εκτός ομάδα από το γενικευμένο γραμμικό Poisson είναι στον παρακάτω πίνακα:

Πίνακας 29: Αναμενόμενα γκολ από το γενικευμένο γρ.μοντέλο Poisson

	λ_x	λ_y
ΑΕΚ - ΟΣΦΠ	0,993501	1,355285
ΠΑΟΚ - ΠΑΟ	1,2625	0,801161
Γιάννενα - ΑΡΗΣ	0,83659	0,849761

Υπολογίζεται ο αριθμός των αναμενόμενων γκολ να είναι μέχρι πέντε γκολ. Είναι ασυνήθιστο ένα παιχνίδι να έχει περισσότερα από πέντε γκολ για μια ομάδα και άρα η πιθανότητα να συμβεί αυτό είναι μικρή. Αυτό θα γίνει δύο φορές μια για την εντός ομάδα και μια για την εκτός ομάδα.

Πίνακας 30: Πιθανότητα η εντός έδρας ομάδα να βάλει 0, 1, 2, 3, 4, 5 γκολ.

	0	1	2	3	4	5
ΑΕΚ - ΟΣΦΠ	37,03%	36,79%	18,27%	6,05%	1,5%	0,3%
ΠΑΟΚ - ΠΑΟ	28,29%	35,72%	22,55%	9,49%	3%	0,76%
Γιάννενα - ΑΡΗΣ	43,32%	36,24%	15,16%	4,23%	0,88%	0,15%

Οι πιθανότητες υπολογίστηκαν με την κατανομή Poisson P(X= x) = $\frac{(\lambda_x)^x \cdot e^{-\lambda x}}{x!}$. Άρα υπάρχει 37,03% πιθανότητα η εντός ομάδα ΑΕΚ να βάλει 0 γκολ εναντίον του Ολυμπιακού. Όμοια υπολογίζονται οι πιθανότητες επί τοις εκατό να βάλουν από μηδέν έως πέντε γκολ οι εντός ομάδες. Αν αθροιστεί η πρώτη σειρά θα βρεθεί 99.94% άρα 0,06 % δεν καλύπτεται με αυτά τα 5 γκολ και είναι πολύ ασήμαντο ποσό.

Όμοια για την εκτός ομάδα

Πίνακας 31: Πιθανότητα η εκτός έδρας ομάδα να βάλει 0, 1, 2, 3, 4, 5 γκολ

	0	1	2	3	4	5
ΑΕΚ - ΟΣΦΠ	25,79%	34,95%	23,68%	10,7%	3,63%	0,98%
ΠΑΟΚ - ΠΑΟ	44,88%	35,96%	14,4%	3,85%	0,77%	0,12%
Γιάννενα - ΑΡΗΣ	42,75%	36,33%	15,44%	4,37%	0,93%	0,16%

Ο Ολυμπιακός έχει 25,79% πιθανότητα να βάλει 0 γκολ με την ΑΕΚ.

Στην συνέχεια υπολογίζονται όλα τα εν δυνάμει συνδυασμένα αποτελέσματα 1-0, 2-0, ..., 5-4, 0-1, 0-2, ..., 4-5, 0-0, ... , 5-5 όπου η εντός ομάδα και η εκτός ομάδα θα βάλουν μέχρι πέντε γκολ. Από το ανεξάρτητο Poisson μοντέλο υπολογίζεται η πιθανότητα για κάθε αποτέλεσμα και κάθε παιχνίδι. Για παράδειγμα $P(X=0, Y=0) = P(X=0)\cdot P(Y=0) = 37,03\%\cdot 25,79\% = 9,55\%$

Πίνακας 32: Πιθανότητες με τα όλα τα δυνατά αποτελέσματα (νίκης-ήττας-ισοαπλίας αντίστοιχα)

	1 - 0	....	5 - 4	0 - 1	....	4 - 5	0 - 0	....	5 - 5
ΑΕΚ - ΟΣΦΠ	9,49%	....	0,01%	12,94 %	....	0,01%	9,55%	....	0%
ΠΑΟΚ - ΠΑΟ	7,16%	....	0,07%	7,29%	....	0,07%	4,74%	....	0,02%
Γιάννενα - ΑΡΗΣ	17,63%	....	0%	9,38%	....	0 %	13,48%	....	0%

Οι πραγματικές πιθανότητες επί τις εκατό εντός νίκης υπολογίζονται αρκεί να προστεθούν όλες οι πιθανότητες με εντός νίκη από 1-0 μέχρι 5-4. Η πιθανότητα να κερδίσει η ΑΕΚ τον Ολυμπιακό είναι $P(X > Y) = P(X=1, Y=0) +...+ P(X=5, Y=4) = 9,49\% + ... + 0,01\% = 28,4\%$. Το ίδιο για τις ισοπαλίες που είναι 0-0 μέχρι 5-5, και για την εκτός νίκη που είναι το άθροισμα όλων των εκτός νικών. Ο παρακάτω πίνακας έχει την πιθανότητα επί τοις εκατό για κάθε αποτέλεσμα.

Πίνακας 33: Πιθανότητα επί τοις εκατό για κάθε αποτέλεσμα

	Νίκη Εντός	Ισοπαλία	Νίκη Εκτός
ΑΕΚ - ΟΣΦΠ	28,4%	27,44%	44,81%
ΠΑΟΚ - ΠΑΟ	48,02%	29,18%	23,58%
Γιάννενα - ΑΡΗΣ	32,5%	34,22%	33,24%

Για να μετατραπούν οι πιθανότητες σε αποδόσεις, διαιρείται το 1 με τις αντίστοιχες πιθανότητες. Η απόδοση να νικήσει η ΑΕΚ τον Ολυμπιακό είναι $\frac{1}{28,4\%} = 3,52$. Άρα υπολογίζονται οι πραγματικές αποδόσεις για κάθε αποτέλεσμα σύμφωνα με το μοντέλο.

Πίνακας 34: Αποδόσεις για κάθε αποτέλεσμα με το glm Poisson

	Νίκη Εντός	Ισοπαλία	Νίκη Εκτός
ΑΕΚ - ΟΣΦΠ	3,52	3,64	2,23
ΠΑΟΚ - ΠΑΟ	2,08	3,43	4,24
Γιάννενα - ΑΡΗΣ	3,08	2,92	3,01

Από τους παραπάνω πίνακες φαίνεται η αδυναμία του μοντέλου να αποτυπώσει σωστά τις ισοπαλίες, διότι στην πραγματικότητα οι ισοπαλίες είναι ένα πιο συχνό φαινόμενο. Ακόμα και στην περίπτωση που οι ομάδες αναμένεται να σκοράρουν τον ίδιο αριθμό γκολ, όπως στον αγώνα Γιάννενα - ΑΡΗΣ, το μοντέλο δεν καταφέρνει να δώσει την υψηλότερη πιθανότητα για ισοπαλία.

5.8 Συμπέρασμα

Χρησιμοποιώντας δεδομένα από αγώνες του ελληνικού πρωταθλήματος, μελετήθηκε το απλό Poisson μοντέλο, το οποίο θεωρεί το λ ως τον μέσο όρο των γκολ, όπου οι παρατηρούμενες με τις αναμενόμενες συχνότητες φαίνεται να είναι κοντά μεταξύ τους. Ωστόσο το τεστ καλής προσαρμογής δείχνει ότι δεν προσαρμόζεται ικανοποιητικά και άρα δεν είναι κατάλληλο από στατιστική άποψη.

Όσον αφορά τα δύο γενικευμένα γραμμικά μοντέλα του Maher ένα τεστ καλής προσαρμογής θα έδειχνε ότι ταιριάζουν καλύτερα και, σε αντίθεση με το παραπάνω απλό Poisson μοντέλο, δεν απορρίπτονται.

Glm Poisson

$$\log(\lambda_x) = -1.903 - 0,035 \cdot \text{επίδραση έδρας} + 1,168 \cdot \text{επίθεση}_x + 0,910 \cdot \text{άμυνα}_y$$

$$\log(\lambda_y) = -2,201 + 1,056 \cdot \text{επίθεση}_y + 1,094 \cdot \text{άμυνα}_x$$

Glm αρνητική διωνυμική

$$\log(\lambda_x) = -2{,}007 - 0{,}055 \cdot \text{επίδραση έδρας} + 1{,}224 \cdot \text{επίθεση}_x + 0{,}962 \cdot \text{άμυνα}_y$$

$$\log(\lambda_y) = -2{,}293 + 1{,}079 \cdot \text{επίθεση}_y + 1{,}159 \cdot \text{άμυνα}_x$$

Αν στο τέλος της σεζόν 2021-2022 η ΑΕΚ παίξει στην έδρα της με αντίπαλο τον Ολυμπιακό, τότε η εκτίμηση των μοντέλων δίνει:

Πίνακας 35: Αναμενόμενα γκολ με Poisson και αρνητική διωνυμική

	Αναμενόμενα γκολ ΑΕΚ	Αναμενόμενα γκολ ΟΣΦΠ
glm Poisson	0,9935	1,3552
glm bn	0,9784	1,3936

Οι δύο κατανομές δίνουν παρόμοια αποτελέσματα, διότι το λ μεταβάλλεται από ένα σχετικά μικρό ποσό από παιχνίδι σε παιχνίδι και άρα η Poisson προσαρμογή δεν διαφέρει από την αρνητική διωνυμική. Αυτό συμβαίνει επίσης εξαιτίας του ότι το δείγμα ήταν πολύ μικρό, και δεν είναι επαρκές να απεικονίσει την ανακρίβεια της Poisson. Προκειμένου να φανεί η διαφορά των δύο κατανομών στα αποτελέσματα θα πρέπει να υπάρχουν περισσότερες παρατηρήσεις και τα παιχνίδια να έχουν μεγαλύτερη διακύμανση στο ποσοστό σκοραρίσματος. Για παράδειγμα αν τα δεδομένα ήταν πολλών χρόνων και ο ρυθμός σκοραρίσματος αλλάζει σημαντικά με τα χρόνια για κάθε ομάδα τότε η διαφορά θα ήταν εμφανής, και τότε το πιο σύνθετο μοντέλο της αρνητικής διωνυμικής θα ήταν προτιμότερο. Άρα οι δύο κατανομές δεν παρουσιάζουν σημαντικές διαφορές και είναι σωστό και αληθοφανές να χρησιμοποιείται το Poisson γενικευμένο γραμμικό μοντέλο για να εκτιμήσει τα γκολ. Η εφαρμογή σε πραγματικά δεδομένα που έγινε συμφωνεί με την παραπάνω δήλωση και την έρευνα.

Το ανεξάρτητο μοντέλο παλινδρόμησης Poisson δεν θεωρείται από τα καλύτερα μοντέλα πρόβλεψης αποτελεσμάτων αγώνων ποδοσφαίρου. Η μέθοδος δεν κάνει αρκετά καλές προβλέψεις, κυρίως διότι δεν προβλέπει ικανοποιητικά τις ισοπαλίες. Βελτιώσεις του μοντέλου έχουν να κάνουν με την μοντελοποίηση της εξάρτησης μεταξύ των πιθανοτήτων για τον αριθμό των γκολ που είναι μικρότερα του δύο, καθώς και με την χρησιμοποίηση δεδομένων από το πιο πρόσφατο παρελθόν.

Η μοντελοποίηση αγώνων ποδοσφαίρου είναι κάτι πολύ δύσκολο διότι στην διαδιακασία εμπλέκεται πολύ τυχαιότητα. Ωστόσο επιχειρήθηκε να εξαχθεί όσο το δυνατόν περισσότερη πληροφορία από δεδομένα με το τελικό αποτέλεσμα παλαιότερων αγώνων. Τα μοντέλα που

μελετήθηκαν προκύπτουν καθαρά από δεδομένα που έχουν να κάνουν με το αποτέλεσμα του αγώνα, άρα δεν λαμβάνουν υπόψη παράγοντες όπως ο καιρός, τραυματισμούς, διοικητικές αλλαγές και άλλες εξωτερικές επιρροές. Όσον αφορά παρόμοια μοντέλα, το καλύτερο μοντέλο που υπάρχει σήμερα είναι ικανό να προβλέψει με επιτυχία περίπου το 53% ενός συνόλου αγώνων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Αθανασιάδης, Η. (1995). *Παραγοντική Ανάλυση Αντιστοιχιών και Ιεραρχική Ταξινόμηση*, Αθήνα: Εκδόσεις ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ.

Δερμάνης, Α. (1986), *Συνορθώσεις Παρατηρήσεων και Θεωρία Εκτίμησης, Τόμος 1*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Δημητράκος, Θ. (2010). *Πιθανότητες*. Πανεπιστήμιο Αιγαίου, Σάμος, Ελλάδα.

Καραπιστόλης, Δ. (1999). *Ανάλυση Δεδομένων και Έρευνα Αγοράς*. Θεσσαλονίκη: ΑΝΙΚΟΥΛΑ.

Κικίλιας, Π. et al. (2001), *Στατιστική – Πιθανότητες*, Αθήνα: Δήρος.

Μενεξές, Γ. (2007), ‘ Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων’, PhD thesis, Τμήμα Εφαρμοσμένης Πληροφορικής, Πανεπιστημίο Μακεδονίας.

Μπεχράκης, Θ. (1999). *Πολυδιάστατη Ανάλυση Δεδομένων: Μέθοδοι και Εφαρμογές*, Αθήνα: ΝΕΑ ΣΥΝΟΡΑ-Α. Α. ΛΙΒΑΝΗΣ.

Νταϊλιάνας, Χ. (2012), ‘ Γενικευμένα Γραμμικά Μοντέλα με Χρήση του Στατιστικού Πακέτου R’, PhD thesis, Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, Εθνικό Μετσόβιο Πολυτεχνείο.

Ντζούφρας Ιωάννης, << Football Analytics: Προβλήματα, μέθοδοι και διασκεδαστική στατιστική>> (εισήγηση στις επιστημονικές διαλέξεις με τίτλο Hub Science, BODOSSAKI LECTURES ON DEMAND, 6 Φεβρουαρίου 2019).

Παπαδημητρίου, Γ. (2004). *Πολυμεταβλητή Στατιστική Ανάλυση: Πανεπιστημιακές Παραδόσεις*. Θεσσαλονίκη: Έκδοση Πανεπιστημίου Μακεδονίας Οικονομικών και Κοινωνικών Επίστημών.

Παπαδόπουλος, Γ. (2015), *Εισαγωγή στις Πιθανότητες και τη Στατιστική*, Αθήνα: Gutenberg.

Ξενόγλωσση

Baio, G., and Blangiardo,M. (2010), ‘Bayesian hierarchical model for the prediction of football results’, *Journal of Applied Statistics*, 37 (2):253-264.

Baker, R. D., and McHale, I.G. (2015), ‘Time varying ratings in association football: the all-time greatest team is..’ , *Journal of the Royal Statistical Society: Series A*, 178(2), 481-492.

Benzecri, J.-P. (1991), ‘ Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data’, .*Journal of the American Statistical Association*, 86(416) : 1112-115.

Benzecri J.-P. & Collaborateurs (1973). *L' Analyse des Données. Vol. 1: Taxinomie. Vol. 2: Analyse des Correspondances*. Paris: Dunod.

Bock, D.E., and Velleman, P.F., and De Veaux, R. (2007), *Stats: Modeling the World*, Boston: Pearson Addison- Wesley.

Breiman, L. (2001), 'Statistical Modeling: The Two Cultures', *Statistical Science*, 16(3): 199-231.

Boshnakov, G., and Kharrat, T., and Mchale I. (2017), 'A bivariate Weibull count model for forecasting association football scores', *International Journal of Forecasting*, 33(2): 458-466

Clarke, S. R., and Norman, J.M. (1995), 'Home Ground Advantage of Individual Clubs in English Soccer', *Journal of the Royal Statistical Society*, 44(4): 509-521.

Clausen, S.-E. (1998). *Applied Correspondence Analysis: An Introduction*. Thousand Oakes: Sage Publications.

Davenport, T. (2014), ' Analytics in Sport : The New Science of Winning ', *INTERNATIONAL ISTITUTE FOR ANALYTICS*, February, p.2.

De Leeuw, J. (2005), ' Multivariate Analysis With Optimal Scaling'. *Department of Statistics, UCLA*. Department of Statistics Papers. Paper 2005103002. Διαθέσιμο στην ιστοσελίδα: http://repositories.cdlib.org/uclastat/papers/2005103002

Dillon, W & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. New York: John Willey & Sons, Inc.

Dixon, M.J., and Coles, S.G. (1997), 'Modelling Association Football Scores and Inefficiencies in the Football Betting Market', *Journal of the Royal Statistical Society*, 46 (2): 265-280.

Dobson, J.A. (2002), *An introduction to generalized linear models*, 2nd edition, Boca Raton: CHAPMAN & HALL/CRC

Goddard, J. (2005), 'Regression models for forecasting goals and results in professional football', *International Journal of Forecasting*, 21(2): 331-340.

Greenwood, P. E., and Nikulin, M.S. (1996), *A Guide to Chi-Squared Testing*, New York: Wiley.

Hilbe, J. (2011), *Negative Binomial Regression*, 2nd edition, New York: Cambridge University Press.

Karlis, D. and Ntzoufras, I. (2003), ' Analysis of sports data using bivariate Poisson models', *Journal of Royal Statistical Society: Series D (The statistician)*, 52(3): 381-393.

Karlis, D. and Ntzoufras, I. (2008), ‘ Bayesian modeling of football outcomes: using the Skellam’s distribution for the goal difference, *IMA Journal of Management Mathematics*, 22(2): 133-145.

Karlis, D. , and Ntzoufras, I. (2000), ‘ On Modeling Soccer Data’, *Student*, 3(4): 229-244.

Liden, J. (2016), ‘Bivariate Models to Predict Football Results’ , PhD thesis, Department of Mathematics, Uppsala University

Maher, M. J. (1982), ‘Modeling association football scores’, *Statistica Neerlandica*, 36 (3): 109-118.

McHale, I.G., and Scarf, P. A. (2007), ‘Modelling soccer matches using bivariate discrete distributions’ , *Statistica Neerlandica*, 61(4): 432-445.

Moroney, M. J. (1956), *Facts from Figures*, 3rd edition, London: Penguin Books.

Pfeiffer, P. (1978), *Concepts of Probability Theory*. New York: Dover Publications, Inc.

Pollard, R. (1985), ‘Goal-Scoring and the Negative Binomial Distribution’, *The Mathematical Gazette*, 69(447):45-47.

Quantitative Psychology. (2018), *Understanding Generalized Linear Models*, accessed at 5 June 2022, available from < https://www.youtube.com/watch?v=SqN-qlQOM5A> .

Reep, C., and Polland, R., and Benjamin, B. (1971), ‘Skill and Chance in Ball Games’, *Journal of the Royal Statistical Society: Series A*, 133(4): 623-629.

Saraiva, E. F. et al. (2016), ‘Predicting football scores via Poisson regression model: applications to the National Football League’, *Communications for Statistical Applications and Methods*, 23 (4): 297-319.

Sclar, A. (1973), ‘Random variables, joint distribution functions, and copulas’, *Kybernetika*, 9(6): 449-460.

Sybil, P. (1994), *McGraw-Hill Dictionary of Scientific and Technical Terms*, New York: McGraw- Hill Companies.

Ziliak, S.T., and McCloske, D.N.(2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, : Michigan, University of Michigan Press.