ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Π.Μ.Σ. «ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ»

ΘΕΜΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

«Αμεροληψία ή μη του Συντελεστή Μεταβλητότητας και άλλες Ιδιότητές του. Η περίπτωση των Διακριτών Μεταβλητών»

ΕΠΙΜΕΛΕΙΑ

ΕΛΕΥΘΕΡΙΑΔΗΣ ΒΑΣΙΛΕΙΟΣ

Α.Μ. 646

Επιβλέπων Καθηγητής	Φαρμάκης Νικόλαος
	Αναπληρωτής Καθηγητής
Συμβουλευτική Επιτροπή	Κολυβά-Μαχαίρα Φωτεινή
	Αναπληρώτρια Καθηγήτρια.
	Παπαδοπούλου Αλεξάνδρα
	Επίκουρη Καθηγήτρια

ΘΕΣΣΑΛΟΝΙΚΗ 19 ΔΕΚΕΜΒΡΙΟΥ 2017

ARISTOTLE UNIVERSITY OF THESSALONIKI

MSc «STATISTICS AND MODELLING»

SUBJECT

«Unbiased or not Coefficient of Variation and its other Properties. The case of Discrete Variables»

WRITER

ELEFTHERIADIS VASILEIOS

R.N. 646

Supervisor Professor	Farmakis Nikolaos	
		Associate Professor
Advisory Committee	Kolyva – Mahaira Fotini	
		Associate Professor
	Papadopoulou Alexandra	
		Assistant Professor

THESSALONIKI 19 DECEMBER 2017

ΠΡΟΛΟΓΟΣ-ΕΥΧΑΡΙΣΤΙΕΣ

Η επιλογή του συγκεκριμένου θέματος έγινε κατόπιν συζήτησης με τον επιβλέποντα καθηγητή μου Νικόλαο Φαρμάκη που μελετάει συστηματικά και αρκετά χρόνια τον τομέα της Δειγματοληψίας. Η εκτίμηση του μεγέθους του δείγματος για την επίτευξη μικρού σφάλματος στο δειγματικό συντελεστή μεταβλητότητας είναι ένα πρόβλημα που δεν έχει μελετηθεί διεξοδικά για αυτό και θεωρήθηκε άξιο μελέτης. Σε αυτή τη μελέτη βοήθησε ο επιβλέπων καθηγητής μου με τη σωστή καθοδήγηση του και την ορθή τοποθέτηση του ερευνητικού ερωτήματος. Ιδιαίτερες ευχαριστίες στην οικογένεια μου για την συνεχή στήριξη τους, στην συμβουλευτική επιτροπή για τις υποδείξεις τους και στον πολύ καλό φίλο και συνάδελφο Χρήστο Πανώρια για τις συμβουλές του στο προγραμματιστικό περιβάλλον της R.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ-ΓΡΑΦΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

CV	συντελεστής μεταβλητότητας πληθυσμού
CVdeig.	δειγματικός συντελεστής μεταβλητότητας
δ.ε.	διάστημα εμπιστοσύνης
Ε.Μ.Π.	εκτιμητής μέγιστης πιθανοφάνειας
MLE(CV)	ο συντελεστής μεταβλητότητας υπολογισμένος με τη μέθοδο της πιθανοφάνειας
NCV	βέλτιστο δείγμα για προσέγγιση του συντελεστή μεταβλητότητας του πληθυσμού από τον δειγματικό συντελεστή μεταβλητότητας
NMLE(CV)	βέλτιστο δείγμα για προσέγγιση του συντελεστή μεταβλητότητας του πληθυσμού από τον συντελεστή μεταβλητότητας υπολογισμένος με τη μέθοδο της πιθανοφάνειας
MSE(CV)	το μέσο τετραγωνικό σφάλμα της εκτίμησης του συντελεστή μεταβλητότητας του πληθυσμού από τον δειγματικό συντελεστή μεταβλητότητας
MSE(MLE(CV))	το μέσο τετραγωνικό σφάλμα της εκτίμησης του συντελεστή μεταβλητότητας του πληθυσμού από τον συντελεστή μεταβλητότητας υπολογισμένος με τη μέθοδο της πιθανοφάνειας

ΠΕΡΙΛΗΨΗ

Η παρούσα Διπλωματική εργασία «Αμεροληψία ή μη του Συντελεστή Μεταβλητότητας και άλλες Ιδιότητές του. Η περίπτωση των Διακριτών Μεταβλητών» μελετάει το μέγεθος του δείγματος που πρέπει να παρθεί στις κυριότερες διακριτές κατανομές έτσι ώστε να έχουμε την αμεροληψία του συντελεστή μεταβλητότητας ή τουλάχιστο μια «μικρή» μεροληψία του. Θέλουμε δηλαδή να βρούμε το μέγεθος του δείγματος που πρέπει να πάρουμε έτσι ώστε ο συντελεστής μεταβλητότητας του δείγματος να είναι ένα αντιπροσωπευτικό μέτρο για τον αντίστοιχο του πληθυσμού. Αυτός είναι και ο βασικός σκοπός του κλάδου της Δειγματοληψίας. Μας ενδιαφέρουν «καλά» αποτελέσματα είτε παίρνουμε δείγματα όπου ένα ή περισσότερα στοιχεία είναι δυνατόν να επαναλαμβάνονται (δειγματοληψία με επανάθεση), είτε δείγματα όπου όλα τα στοιχεία είναι διαφορετικά μεταξύ τους (δειγματοληψία χωρίς επανάθεση). Στόχος είναι για δεδομένο απόλυτο σφάλμα προσέγγισης (που επιλέγει κάθε φορά ο χρήστης) του θεωρητικού συντελεστή μεταβλητότητας από τον δειγματικό ,να βρίσκεται κάθε φορά το κατάλληλο μέγεθος του δείγματος. Η μέθοδος στηρίχτηκε σε γνώσεις της Στατιστικής, της Θεωρίας Πιθανοτήτων, της Εκτιμητικής αλλά και σε γνώσεις του προγραμματιστικού περιβάλλοντος της R. Πολύ σημαντική ήταν επίσης η μέθοδος της μέγιστης Πιθανοφάνειας. Τα αποτελέσματα ήταν πολύ ικανοποιητικά καθώς επαληθεύτηκαν οι χρήσιμες ιδιότητες των Ε.Μ.Π.(εκτιμητές μέγιστης πιθανοφάνειας). Επιπλέον για δεδομένο απόλυτο σφάλμα στη δειγματοληψία με επανάθεση βρίσκεται πάντα το κατάλληλο μέγεθος δείγματος, ενώ στη δειγματοληψία χωρίς επανάθεση ένα πολύ «μικρό» δείγμα μπορεί να είναι αρκετά αξιόπιστο.

ABSTRACT

The current master thesis «Unbiased or not Coefficient of Variation and its other Properties. The case of Discrete Variables» studies the sample size should be taken in the main discrete distributions so that we succeed the impartiality of the coefficient of variation or at least its "small" bias. We want to find the sample number we have to take so that the coefficient of variation of the sample is a representative measure for the corresponding population. This is also the main purpose of the Sampling sector. We are interested in "good" results either we take samples where one or more elements can be repeated (sampling with replay), or samples where all elements are different from each other (sampling without repositioning). The main purpose is to find the appropriate size of the sample for a given absolute approximation error (chosen by the user each time) of the theoretical coefficient of variation from its sample. The method was based on knowledge of Statistics ,Theory of Probabilities, Estimation and knowledge of the programming environment in R. Very important was the method of maximum likelihood estimation. The results were very satisfactory because the useful qualities of the MLE (maximum likelihood estimators) were confirmed. Furthermore, for a given absolute error in sampling with replay we can always find the appropriate sample size, while in sampling without repositioning a very "small" sample can be quite reliable.

ΕΙΣΑΓΩΓΗ

Η παρούσα διπλωματική εργασία μελετάει την εκτίμηση του μεγέθους του δείγματος για τις κυριότερες διακριτές κατανομές, έτσι ώστε να επιτευχθεί η αμεροληψία του δειγματικού συντελεστή μεταβλητότητας. Σε πρόσφατες μελέτες έχει αναφερθεί τύπος για αμερόληπτο εκτιμητή αλλά μόνο την περίπτωση της κανονικής κατανομής (Herve Abdi 2010) ενώ εκτιμήθηκε και ο αμερόληπτος εκτιμητής για το τετράγωνο του συντελεστή μεταβλητότητας υπό προϋποθέσεις (Robert Breunig 2001).

Στόχος της εργασίας αυτής είναι για δεδομένο απόλυτο σφάλμα μεταξύ του θεωρητικού συντελεστή μεταβλητότητας και του δειγματικού, να γίνει η εκτίμηση του κατάλληλου μεγέθους δείγματος, σε μορφή διαστήματος εμπιστοσύνης. Έτσι θα αναπτυχθεί ο κλάδος της Δειγματοληψίας στην περίπτωση των διακριτών κατανομών. Ανάλογη μελέτη μπορεί να γίνει και για τις συνεχείς κατανομές. Ο προβληματισμός που οδήγησε στη συγκεκριμένη έρευνα ήταν ότι για μικρά δείγματα παρατηρήθηκε πολύ κακή προσέγγιση του θεωρητικού συντελεστή μεταβλητότητας από το δειγματικό.

Η μεθοδολογία που ακολουθήθηκε στηρίχτηκε σε γνώσεις του κλάδου της Στατιστικής και των Πιθανοτήτων. Χρησιμοποιήθηκε η μέση τιμή και η τυπική απόκλιση των κυριότερων διακριτών κατανομών και έγινε και χρήση της μεθόδου της μέγιστης πιθανοφάνειας για τον υπολογισμό παραμέτρων των κατανομών από διάφορα δείγματα. Υπολογίστηκε ο θεωρητικός συντελεστής μεταβλητότητας και έπειτα ο δειγματικός με ή χωρίς την μέθοδο της μέγιστης πιθανοφάνειας. Έπειτα στο προγραμματιστικό περιβάλλον της R,για δεδομένο απόλυτο σφάλμα που επιλέγει ο χρήστης μεταξύ θεωρητικού συντελεστή μεταβλητότητας και δειγματικού, εκτιμήθηκε το μέγεθος του δείγματος, είτε με την μέθοδο της μέγιστης πιθανοφάνειας είτε χωρίς.

Σε μια σύντομη παρουσίαση της διάθρωσης των κεφαλαίων αναφέρουμε ότι στο Κεφάλαιο 1, υπολογίζεται ο θεωρητικός συντελεστής μεταβλητότητας για τις κυριότερες διακριτές κατανομές, στο Κεφάλαιο 2 αναπτύσσονται οι βασικές γνώσεις της Εκτιμητικής, στο Κεφάλαιο 3 υπολογίζονται οι συντελεστές μεταβλητότητας με τη μέθοδο της Πιθανοφάνειας ενώ στα Κεφάλαια 4 και 5 υλοποιείται η θεωρία των προηγούμενων κεφαλαίων στο προγραμματιστικό περιβάλλον της R.

ΚΕΦΑΛΑΙΟ 1 «ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ»

ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό θα αναφερθούμε στις εισαγωγικές έννοιες της Στατιστικής και της Θεωρίας Πιθανοτήτων που θεωρούνται απαραίτητες προκειμένου να γίνει η κατανόηση της παρούσας μελέτης.

Θα περιγραφούν τα αριθμητικά μέτρα που χρησιμοποιήθηκαν στη συγκεκριμένη μελέτη. Επίσης θα γίνει αναφορά στον τρόπο που προκύπτουν τα αριθμητικά μέτρα διακριτών κατανομών μέσω της Θεωρίας Πιθανοτήτων.

Πιο συγκεκριμένα θα μελετηθεί η μέση τιμή, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας. Θα αναφερθούν οι ορισμοί και οι τύποι των παραπάνω αριθμητικών μέτρων, καθώς επίσης και πως προκύπτουν με χρήση της συνάρτησης πιθανότητας. Ακόμη, θα δώσουμε τον ορισμό των πιθανογεννητριών και θα μελετήσουμε κάποιες ιδιότητες τους. Τέλος θα υπολογιστούν τα παραπάνω αριθμητικά μέτρα για τις κυριότερες διακριτές κατανομές.

1.1 ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ ΣΤΑΤΙΣΤΙΚΗΣ

Πριν αναφερθούμε στα αριθμητικά περιγραφικά μέτρα θα αναφερθούμε στις έννοιες της Περιγραφικής στατιστικής, του πληθυσμού και του δείγματος.

Ορισμός 1.1 **Περιγραφική** στατιστική είναι ο κλάδος της στατιστικής που ασχολείται με την οργάνωση, συγκέντρωση και περιγραφή ενός **συνόλου δεδομένων**. Το σύνολο δεδομένων που μας ενδιαφέρει αφορά τα στοιχεία (ή άτομα) ενός συνόλου αναφοράς που λέγεται **πληθυσμός.**

Στην πραγματικότητα είναι (συνήθως) αδύνατο να συγκεντρωθούν όλα τα δεδομένα από τον εκάστοτε πληθυσμό λόγω χρόνου ή κόστους. Για αυτόν τον λόγο επικεντρωνόμαστε σε ένα μέρος του πληθυσμού για να αντλήσουμε πληροφορία για τον πληθυσμό. Το υποσύνολο του Πληθυσμού ονομάζεται **δείγμα.**

Ορισμός 1.2 **Δείγμα** είναι ένα υποσύνολο ατόμων του πληθυσμού παρμένα με τυχαίο τρόπο.

Από τα στοιχεία (άτομα) του δείγματος θα προκύψουν τα δεδομένα μας (δειγματικά δεδομένα). Αυτά μπορεί να είναι ποσοτικά ή ποιοτικά. Συνήθως τα δεδομένα έχουν τη μορφή της μεταβλητής που ονομάζεται τυχαία μεταβλητή (τμ *X, βλέπε σελίδα 21 τον αναλυτικό ορισμό*).

Ορισμός 1.3 Ονομάζουμε **ποιοτική τη μεταβλητή** που τα αντίστοιχα της δεδομένα δεν εκφράζουν κάτι το μετρήσιμο.

Για παράδειγμα σε μία ερώτηση σχετικά με το τι ασχολείται ένα άτομο με τον ελεύθερό του χρόνο θα έχουμε ως απαντήσεις ποιοτικά δεδομένα αφού πιθανές απαντήσεις θα είναι αθλητισμός, χορός, μουσική και διασκέδαση με φίλους. Η τμ *X* εδώ μπορεί να αποδοθεί με το όρο «ασχολία κατά τον ελεύθερο χρόνο». Ανάλογα μπορεί να έχουμε τμ *Y*= «διεύθυνση κατοικίας» με πιθανές τιμές «Μ. Αλεξάνδρου 17», «Βενιζέλου 22», κλπ.

Ορισμός 1.4 Αν τα χαρακτηριστικά μίας μεταβλητής παίρνουν μόνο αριθμητικές τιμές τότε η μεταβλητή ονομάζονται **ποσοτική** τυχαία μεταβλητή.

Για παράδειγμα το βάρος ή το ύψος ενός ατόμου θα μας δώσει ποσοτικά δεδομένα ως τιμές της αντίστοιχης τμ. .

Ωστόσο για την περιγραφή ποιοτικών και ποσοτικών δεδομένων πρέπει να οριστούν οι τιμές (επίπεδα, levels) με τέτοιο τρόπο ώστε κάθε παρατήρηση να αντιστοιχεί σε μία και μόνο τιμή. Παρακάτω ορίζουμε τις έννοιες της συχνότητας και της σχετικής συχνότητας μιας τιμής της τμ *X*.

Ορισμός 1.5 Ονομάζεται **συχνότητα τιμής της τμ *X*** ο αριθμός των παρατηρήσεων που αντιστοιχεί σ' αυτήν την τιμή ενώ **σχετική συχνότητα τιμής** ονομάζεται η αναλογία του πληθυσμού των παρατηρήσεων της συγκεκριμένης τιμής ως προς το συνολικό αριθμό των παρατηρήσεων του δείγματος.

Παρακάτω ο τρόπος παρουσίασης ποσοτικών δεδομένων.

Ορισμός 1.6 Ο πιο συνηθισμένος τρόπος για να περιγράψουμε ποσοτικά δεδομένα του δείγματος είναι το ιστόγραμμα συχνοτήτων ή σχετικών συχνοτήτων που ονομάζεται και **κατανομή σχετικών συχνοτήτων.**

Συνεπώς αφού ολοκληρώσαμε με τη βασική θεωρία των δεδομένων ενός δείγματος προχωράμε στα αριθμητικά περιγραφικά μέτρα που τα περιγράφουν.

Αριθμητικά περιγραφικά μέτρα για μία τμ X είναι οι αριθμοί που υπολογίζονται από το δείγμα ή από τον πληθυσμό και βοηθούν στη δημιουργία μιας εικόνας για την κατανομή. Τα μέτρα αυτά χωρίζονται σε μέτρα κεντρικής τάσης, μέτρα μεταβλητότητας και μέτρα ασυμμετρίας.

Ορισμός 1.7 Δειγματική μέση τιμή ενός συνόλου n μετρήσεων (ή παρατηρήσεων του δείγματος) $x_1, x_2, \ldots\ldots, x_n$ ονομάζεται η τιμή που προκύπτει από τον τύπο: $m= \frac{1}{n}\sum_{i=1}^{n} xi$.

Η δειγματική μέση τιμή ανήκει στα μέτρα κεντρικής τάσης.

Ορισμός 1.8 Εύρος δείγματος ονομάζεται η διαφορά ανάμεσα στην μεγαλύτερη και στην μικρότερη τιμή του. Αντίστοιχα ορίζεται και το εύρος του Πληθυσμού.

Ορισμός 1.9 Διασπορά δείγματος ενός συνόλου n μετρήσεων $x_1, x_2, \ldots\ldots, x_n$ ονομάζεται η τιμή που προκύπτει από τον τύπο:

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(xi - m)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} xi^2 - n\, m^2\right)$.

Ορισμός 1.10 Τυπική απόκλιση δείγματος ονομάζεται η θετική τετραγωνική ρίζα της δειγματικής διασποράς και δίνεται από τον τύπο $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(xi - m)^2}$.

Το εύρος δείγματος, η διασπορά και η τυπική απόκλιση δείγματος ανήκουν στα μέτρα μεταβλητότητας. Αναφέρονται στο πόσο ‘απλώνεται’ η κατανομή του πληθυσμού. Το μειονέκτημα του εύρους είναι ότι δεν έχουμε πληροφορία για το πώς απλώνονται οι τιμές του γύρω από την μέση τιμή αλλά έχουμε πληροφορία μόνο για το ποια είναι η διαφορά των ακραίων τιμών. Αυτό το μειονέκτημα προσπερνιέται με χρήση της διασποράς και της τυπικής απόκλισης. Ωστόσο επειδή η διασπορά δε μετριέται στις μονάδες μέτρησης των τιμών αλλά με τα τετράγωνα των μονάδων μέτρησης, προτιμούμε την τυπική απόκλιση που μετριέται στις μονάδες μέτρησης των τιμών. Η τυπική απόκλιση παρόλο που είναι ένα αξιόπιστο μέτρο μεταβλητότητας δεν μπορεί να μας δώσει καλή εικόνα για την σύγκριση της μεταβλητότητα δύο δειγμάτων όταν αυτές μετριούνται σε διαφορετική μονάδα μέτρησης ή όταν οι μέσες τιμές διαφέρουν πολύ.

Οπότε δημιουργείται η ανάγκη για ένα μέτρο απαλλαγμένο από τις μονάδες μέτρησης και που δε θα απαιτεί παρόμοιες μέσες τιμές μεταξύ των δειγμάτων για να συγκρίνουμε την μεταβλητότητα. Αυτό το μέτρο είναι ο συντελεστής μεταβλητότητας.

Ορισμός 1.11 Συντελεστής μεταβλητότητας λέγεται ο λόγος της δειγματικής τυπικής απόκλισης προς το δειγματικό μέσο, δηλαδή CV= $\frac{S}{m}$, ή CV= $\frac{S}{m}100\%$. Ένα δείγμα θεωρείται ομοιογενές ως προς την τμ *X* (δηλαδή οι τιμές της *X* διαφέρουν μεταξύ τους ελάχιστα, μικρή διασπορά της τμ *X*) όταν η τιμή του CV είναι το πολύ 10%.

Από τον τύπο φαίνεται ότι επειδή η τυπική απόκλιση και η μέση τιμή μετριούνται με ίδιες μονάδες ο συντελεστής μεταβλητότητας δεν εξαρτάται από την μονάδα μέτρησης (αδιάστατο μέγεθος, unit free quantity). Επίσης για να κρίνουμε την μεταβλητότητα δε παίζει ρόλο μόνο η τυπική απόκλιση αλλά και το μέτρο κεντρικής τάσης της μέσης τιμής.

Για παράδειγμα έστω ότι έχουμε την επίδοση φοιτητών στα μαθήματα της Δειγματοληψίας και της Στατιστικής.

Με άριστα το 100 θεωρούμε ότι πήραμε τα εξής αποτελέσματα:

Για την Δειγματοληψία: $m_{Δειγ}=60$ και $S_{Δειγ}=5$.

Για την Στατιστική: $m_{Στατ}=78$ και $S_{Στατ}=6$.

Τότε $CV_{Δειγ.}=\frac{S_{Δειγ}}{m_{Δειγ}}=\frac{5}{60}=0.083=8.3\%$.

Και $CV_{Στατ}=\frac{S_{Στατ}}{m_{Στατ}}=\frac{6}{78}=0.077=7.7\%$.

Έτσι παρόλο που οι βαθμοί στην Στατιστική είχαν μεγαλύτερη τυπική απόκλιση τελικά εμφανίζουν μικρότερη μεταβλητότητα λόγω της μέσης τιμής της που είναι αρκετά μεγάλη σε σχέση με αυτής της Δειγματοληψίας.

Τέλος παρακάτω ορίζουμε κάποια μέτρα ασυμμετρίας.

Ορισμός 1.12 Η **δειγματική κεντρική ροπή τάξης r**,δίνεται από τη σχέση:

$m_r=\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^r$, r=1,2,…..

Ορισμός 1.13 Η **δειγματική ροπή τάξεως r**, δίνεται από τη σχέση :

$m_r' = \frac{1}{n} \sum_{i-1}^{n} xi^r$, r=1,2,…..

1.2 ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ ΤΗΣ ΘΕΩΡΙΑΣ ΠΙΘΑΝΟΤΗΤΩΝ

Όπως αναφέρθηκε προηγούμενα η μελέτη του πληθυσμού είναι μια ανέφικτη διαδικασία για αυτό μελετάμε ένα μέρος του πληθυσμού, το **δείγμα**. Καταφεύγουμε λοιπόν σε **δειγματοληψία** και, για να είναι αντιπροσωπευτικό το δείγμα, πρέπει να είναι παρμένο με τυχαίο τρόπο. Η δειγματοληψία μπορεί να γίνει **με επανάθεση**, (με δυνατότητα να πάρουμε το ίδιο στοιχεί περισσότερες φορές από μια) είτε **χωρίς επανάθεση**, παίρνοντας δηλαδή μόνο διαφορετικά στοιχεία.

Ορισμός 1.14 Όλα τα δυνατά αποτελέσματα μιας δειγματοληψίας αποτελούν το δειγματοχώρο που συμβολίζεται συνήθως με Ω. Κάθε δυνατό αποτέλεσμα του δειγματοχώρου, δηλαδή κάθε σημείο του δειγματοχώρου λέγεται **απλό γεγονός ή ενδεχόμενο,** ενώ ένα σύνολο απλών γεγονότων **λέγεται σύνθετο γεγονός.** Οι δειγματοχώροι που έχουν πεπερασμένο ή αριθμήσιμο πλήθος στοιχείων λέγονται **διακριτοί**, ενώ αυτοί που έχουν μη αριθμήσιμο πλήθος στοιχείων λέγονται **συνεχείς.**

Παρακάτω ο πίνακας των γεγονότων

Δειγματοχώρος Ω	Σύνολο αναφοράς Ω
Αδύνατο γεγονός	Σύνολο Ø
Απλό γεγονός Α	Σύνολο Α
Δεν συμβαίνει το γεγονός Α	Σύνολο Ω-Α
Τα γεγονότα Α και Β συμβαίνουν ταυτόχρονα	Σύνολο Α∩Β
Τουλάχιστον ένα από τα γεγονότα Α,Β συμβαίνει.	Σύνολο Α∪ Β.

Πίνακας 1: Γεγονότα και Σύνολα

Ορισμός 1.15 Δύο γεγονότα λέγονται Α,Β λέγονται **ασυμβίβαστα ή ξένα** όταν η πραγματοποίηση του ενός γεγονότος αποκλείει την πραγματοποίηση του άλλου, δηλαδή

Α,Β ασυμβίβαστα ↔ Α∩Β= Ø .

Συνεπώς τώρα μπορούμε να προχωρήσουμε στον ορισμό της πιθανότητας.

Ορισμός 1.16 Η πιθανότητα σαν όριο σχετικής συχνότητας :

« Αν στις Ν επαναλήψεις ενός πειράματος ένα γεγονός Α εμφανίστηκε N_A φορές, τότε το πηλίκο $f_A = N_A / N$ ονομάζεται **σχετική συχνότητα** του γεγονότος Α. Όσο το Ν μεγαλώνει τόσο η σχετική συχνότητα σταθεροποιείται γύρω από έναν αριθμό που ονομάζεται πιθανότητα του γεγονότος Α και συμβολίζεται με Ρ(Α)».

Ορισμός 1.17 Αξιωματικός ορισμός της πιθανότητας(Kolmogorov,1930)

Η πιθανότητα είναι μία συνολοσυνάρτηση που ικανοποιεί τα παρακάτω αξιώματα:

i) $P(\Omega) = 1$

ii) $0 \leq P(A) \leq 1, \ \forall A \subseteq \Omega$.

iii) $P(A_1 \cup A_2 \cup \ldots\ldots A_\kappa) = P(A_1) + P(A_2) + \ldots\ldots + P(A_\kappa) \ \forall A_\iota \subseteq \Omega$ και

$A_i \cap A_j = \emptyset$, $\forall i \neq j$ δηλαδή τα γεγονότα $A_1, A_2, \ldots\ldots, A_\kappa$ είναι ανά δύο **ασυμβίβαστα.**

Συνεχίζουμε με τον ορισμό της τυχαίας μεταβλητής.

Ορισμός 1.18 Τυχαία μεταβλητή ονομάζεται μια συνάρτηση που απεικονίζει το σύνολο των δυνατών αποτελεσμάτων μιας δειγματοληψίας στο σύνολο των πραγματικών αριθμών.

Η τυχαία μεταβλητή μπορεί να είναι συνεχής η διακριτή. Να σημειώσουμε εδώ ότι αριθμήσιμο λέγεται ένα σύνολο που έχει πλήθος στοιχείων όσο και το σύνολο των φυσικών αριθμών.

Ορισμός 1.19 Αν η τυχαία μεταβλητή παίρνει πεπερασμένο ή αριθμήσιμο πλήθος τιμών τότε λέγεται **διακριτή** ενώ αν παίρνει τιμές σε ένα διάστημα (α, β) με

$-\infty \leq \alpha < \beta \leq \infty$ θα λέγεται **συνεχής.**

Παρακάτω ορίζουμε την συνάρτηση πιθανότητας, θεωρώντας γνωστή την έννοια της συνάρτησης.

Ορισμός 1.20 Αν η τυχαία μεταβλητή είναι διακριτή, τότε η συνάρτηση που δίνει την πιθανότητα η τυχαία μεταβλητή X να πάρει την τιμή x, λέγεται **συνάρτηση πιθανότητας** της τυχαίας μεταβλητής X, συμβολίζεται με $f_X(x) = P(X=x)$ και έχει τις εξής ιδιότητες :

i) $f_X(x) \geq 0 \ \forall x$,

ii) $\Sigma f_X(x) = 1$ (το x παίρνει όλες τις τιμές της τυχαίας μεταβλητής X)

Παρακάτω ορίζουμε την συνάρτηση αθροιστικής κατανομής

Ορισμός 1.21 Η συνάρτηση $F_X(x)$ που ορίζεται $F_X(x) = P(X \leq x) \ \forall x \in \mathbb{R}$ ονομάζεται **συνάρτηση αθροιστικής κατανομής** της τυχαίας μεταβλητής X και δίνει την πιθανότητα η τυχαία μεταβλητή X να πάρει όλες τις τιμές της μέχρι το σημείο x.

Παρακάτω ορίζουμε της μέση τιμή της τυχαίας μεταβλητής $g(X)$.

Ορισμός 1.22 Μέση τιμή της τυχαίας μεταβλητής $g(X)$ ορίζεται η :

$Eg(X) = \sum_x g(x)f(x)$ όπου X είναι διακριτή τυχαία μεταβλητή και $f(x)$ η συνάρτηση πιθανότητας της τυχαίας μεταβλητής X.

Συνεπώς $\mu = EX = \sum_x xf(x)$.

Ισχύουν οι ιδιότητες :

$E(\alpha\, g(X)) = \alpha\, E(g(X))$ όπου $\alpha \in \mathbb{R}$ και

$E(g(X)+h(X)) = E(g(X) + h(X))$, (και η γενίκευση για n συναρτήσεις με επαγωγή)

$E(g(X)-h(X)) = E(g(X) - h(X))$.

Παρακάτω ορίζουμε την διασπορά και την τυπική απόκλιση της τυχαίας μεταβλητής X.

Ορισμός 1.23 Διασπορά ή διακύμανση της τυχαίας μεταβλητής X, ορίζεται η:

$\sigma^2 = \mathrm{Var}X = E(X-\mu)^2$,όπου $\mu = EX$, ενώ **τυπική απόκλιση** της τυχαίας μεταβλητής X ορίζεται η $\sigma = \sqrt{\mathrm{Var}X} = \sqrt{E(X-\mu)^2}$. Συνεπώς $\sigma = \sqrt{\sum_x (x-\mu)^2 f(x)}$.

Από τα παραπάνω συμπεραίνουμε ότι για τον συντελεστή μεταβλητότητας της τυχαίας μεταβλητής X έχουμε:

CV= $\frac{\sigma}{\mu} = \frac{\sqrt{\sum_x (x-\mu)^2 f(x)}}{\sum_x x f(x)}$ όπου f(x) η συνάρτηση πιθανότητας της τυχαίας μεταβλητής *X*.

Πρόταση 1.1 Ισχύει ότι $\sigma^2 = EX^2 - (EX)^2$.

Απόδειξη

$\sigma^2 = E(X-\mu)^2 = E(X^2 + \mu^2 - 2\mu X) = EX^2 + (EX)^2 - 2\mu EX = EX^2 + (EX)^2 - 2(EX)^2 = EX^2 - (EX)^2$.

Τέλος απόδειξης.

Πρόταση 1.2 Ισχύει η ιδιότητα: $Var(aX+b) = a^2 VarX$.

Απόδειξη

$Var(aX+b) = E[(aX+b)-(a\mu+b)]^2 = E[a(X-\mu)]^2 = E[a^2 (X-\mu)^2] = a^2 E(X-\mu)^2 =$

$a^2 VarX$.

Τέλος απόδειξης.

Παρακάτω ορίζουμε και τις ροπές τυχαίων μεταβλητών.

Ορισμός 1.24 Αν *X* τυχαία μεταβλητή διακριτή , τότε **κεντρική ροπή r τάξης** λέγεται η ποσότητα:

$\mu_r = E[(X-\mu)^r] = \sum_x (x-\mu)^r p(X=x)$, r=1,2,……

με p(*X*=x) η συνάρτηση πιθανότητας της *X*.

Προϋπόθεση ότι $\sum_x |(x-\mu)^r| p(x) < \infty$.

Ορισμός 1.25 Αν *X* τυχαία μεταβλητή διακριτή , τότε **απλή ροπή r τάξης** λέγεται η ποσότητα :

$\mu'_r = EX^r = \sum_x x^r p(X=x)$, r=1,2,……. με p(*X*=x) η συνάρτηση πιθανότητας της *X*.

Ορισμός 1.26 Αν *X* τυχαία μεταβλητή , τότε **παραγοντική ροπή τάξης r** λέγεται η ποσότητα:

E$X(X$-1)…..(X- r+1), r=1,2,…

<u>Ορισμός 1.27</u> Ο **συντελεστής ασυμμετρίας του Fischer** δίνεται από τη σχέση:

$\gamma_1 = \frac{\mu 3}{\sqrt{\mu 2^3}}$

Πριν προχωρήσουμε στη μελέτη των αριθμητικών μέτρων των κυριότερων διακριτών κατανομών θα αναφερθούμε στις Πιθανογεννήτριες και σε κάποιες ιδιότητες τους που θα χρησιμοποιήσουμε παρακάτω.

<u>Ορισμός 1.28</u> Πιθανογεννήτρια της τυχαίας μεταβλητής X, ορίζεται η συνάρτηση

$\Pi(z)= Ez^x$ όπου z πραγματική παράμετρος.

Οι πιθανογεννήτριες χρησιμοποιούνται κυρίως σε διακριτές τυχαίες μεταβλητές που παίρνουν τις ακέραιες τιμές x=0,1,2,……

Θεωρώντας γνωστές τις έννοιες των ακολουθιών, σειρών και της σύγκλισης έχουμε ότι αν p(x) είναι η συνάρτηση πιθανότητας της τυχαίας μεταβλητής X, τότε $\Pi(z)=\sum_{x=0}^{\infty} p(x)z^x$ και η Π(z) είναι μια δυναμοσειρά.

Επειδή $\sum_{x=0}^{\infty} p(x) =1$ και $p(x) \geq 0$,τότε Π(1)=1 ,δηλαδή η δυναμοσειρά συγκλίνει απόλυτα για $-1\leq z \leq 1$.

Πιο συγκεκριμένα $\sum_{x=0}^{\infty} |z^x p(x)| = \sum_{x=0}^{\infty} |z^x| p(x) \leq \sum_{x=0}^{\infty} p(x) =1$

(αφού συγκλίνει απόλυτα συγκλίνει και απλά).

Θεωρώντας γνωστές τις έννοιες της συνέχειας, της παραγώγισης και της ολοκλήρωσης έχουμε ότι από τις δυναμοσειρές η ολοκλήρωση και η παραγώγιση μπορούν να γίνουν κατά όρους, δηλαδή

$\Pi'(z)= \sum_{x=1}^{\infty} xp(x)z^{x-1}$,$-1 < z < 1$.

$\Pi^{(\kappa)}(z)= \sum_{x=1}^{\infty} x(x-1) \dots \dots \dots (x-\kappa+1)p(x)z^{x-\kappa}$,$-1 < z < 1$.

Αν υπάρχει η EX, τότε η $\Pi'(z)$ είναι συνεχής στο $-1\leq z \leq 1$ και EX=$\Pi'(1)$.

Αν EX=∞, τότε $\Pi'(z)\rightarrow \infty$ όταν $z\rightarrow 1$.

Όμοια αν υπάρχει κ-στη παραγοντική ροπή τότε

$E\mathit{X}(\mathit{X}\text{-}1)\ldots\ldots.(\mathit{X}\text{-}\kappa\text{+}1)= \Pi^{(\kappa)}(1)$.

Αν $\Pi^{(\kappa)}(z)) \rightarrow \infty$ όταν $z \rightarrow 1$,τότε η $E\mathit{X}^{\kappa}$ δεν υπάρχει για κ=1,2,….

Τέλος για την διασπορά έχουμε $\sigma^2 = E\mathit{X}^2 - (E\mathit{X})^2 = E\mathit{X}(\mathit{X}\text{-}1) + E\mathit{X} - (E\mathit{X})^2$,δηλαδή

$\sigma^2 = \Pi''(1) + \Pi'(1) - (\Pi'(1))^2$ **(Σχέση 1.1)**

Τα παραπάνω αναφερόταν σε μονοδιάστατες τυχαίες μεταβλητές. Παρακάτω θα ορίσουμε αντίστοιχες έννοιες για πολυδιάστατες τυχαίες μεταβλητές.

Ορισμός 1.29 Έστω $\mathit{X}=(\mathit{X}_1,\mathit{X}_2,\ldots\ldots,\mathit{X}_n)$ πολυδιάστατη τυχαία μεταβλητή διακριτή. Τότε η **συνάρτηση πιθανότητας** θα είναι $P_{\mathit{X}}(x)= P(\mathit{X}_1 = x_1, \mathit{X}_2= x_2,\ldots\ldots..,\mathit{X}_n= x_n)$, όπου $x=(x_1 , x_2,\ldots..,x_n)$.

Ισχύουν οι γνωστές ιδιότητες όπως και στις μονοδιάστατες μεταβλητές, δηλαδή:

$P_{\mathit{X}}(x) \geq 0$ και $\sum_{x\in\mathbb{Z}^n} p(x) = 1$.

Ορισμός 1.30 Έστω $\mathit{X}=(\mathit{X}_1,\mathit{X}_2,\ldots\ldots,\mathit{X}_n)$ πολυδιάστατη τυχαία μεταβλητή διακριτή. Τότε η **συνάρτηση αθροιστικής κατανομής** θα είναι

$F_{\mathit{X}}(x)=P(\mathit{X}_1\leq x_1,\mathit{X}_2\leq x_2,\ldots\ldots..,\mathit{X}_n\leq x_n)=$
$\sum_{X1=\min\{X1\}}^{x1}\sum_{X2=\min\{X2\}}^{x2}\ldots\ldots\ldots.\sum_{Xn=\min\{Xn\}}^{xn} P(x1, x2, \ldots xn)$

όπου $x=(x_1 , x_2,\ldots..,x_n)$.

Στον ορισμό της μέσης τιμής ισχύουν τα ίδια όπως στις μονοδιάστατες μεταβλητές.

Ορισμός 1.31 Έστω $\mathit{X}=(\mathit{X}_1,\mathit{X}_2,\ldots\ldots,\mathit{X}_n)$ πολυδιάστατη τυχαία μεταβλητή διακριτή. Τότε **μέση τιμή** της $g(\mathit{X})=g(\mathit{X}_1,\mathit{X}_2,\ldots\ldots,\mathit{X}_n)$ ορίζεται η

$E(g(\mathit{X}))= \sum_{X1}\sum_{X2}\ldots\ldots\ldots.\sum_{Xn} g(X1, X2, \ldots\ldots, Xn)\ \ p(X1, X2, \ldots\ldots Xn)$.

Παρακάτω ορίζουμε την έννοια των ανεξάρτητων μεταβλητών.

Ορισμός 1.32 Έστω X, Y τυχαίες μεταβλητές διακριτές. Αυτές θα λέγονται **ανεξάρτητες** όταν $p(\mathit{X}=x,\mathit{Y}=y)= p(\mathit{X}=x)\ p(\mathit{Y}=y)\ \forall x, y$.

Παρακάτω ορίζουμε την συνδιασπορά τυχαίων μεταβλητών .

Ορισμός 1.33 Έστω X,Y τυχαίες μεταβλητές διακριτές. **Συνδιασπορά** των τυχαίων μεταβλητών X,Y ορίζουμε την ποσότητα $\mathrm{Cov}(X,Y)=\mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)$.

Πρόταση 1.34 Ισχύει ότι $\mathrm{Cov}(X,Y) = \mathrm{E}XY - \mathrm{E}X\,\mathrm{E}Y$.

Απόδειξη

$\mathrm{Cov}(X,Y)=\mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)=\mathrm{E}(XY-X\mathrm{E}Y-Y\mathrm{E}X+\mathrm{E}X\mathrm{E}Y)=\mathrm{E}XY-\mathrm{E}(X\mathrm{E}Y)-\mathrm{E}(Y\mathrm{E}X)+\mathrm{E}(\mathrm{E}X\mathrm{E}Y)=\mathrm{E}XY-\mathrm{E}X\mathrm{E}Y-\mathrm{E}Y\mathrm{E}X+\mathrm{E}X\mathrm{E}Y=\mathrm{E}XY-\mathrm{E}X\,\mathrm{E}Y$.

Τέλος απόδειξης

Πρόταση 1.4 Ισχύει ότι αν X,Y ανεξάρτητες διακριτές μεταβλητές τότε $\mathrm{E}XY=\mathrm{E}X\,\mathrm{E}Y$.

Απόδειξη

$\mathrm{E}XY=\sum_X\sum_Y XY\ \mathrm{p}(X,Y)=\sum_X\sum_Y XY\ \mathrm{p}(X)\mathrm{p}(Y)=\sum_X X\mathrm{p}(X)\sum_Y Y\mathrm{p}(Y)=\mathrm{E}X\,\mathrm{E}Y$.

Τέλος απόδειξης

Πόρισμα 1.1 Αν X,Y ανεξάρτητες διακριτές τότε $\mathrm{Cov}(X,Y)=0$.

Πρόταση 1.5 Ισχύει ότι $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$.

Απόδειξη

$\mathrm{Var}(X+Y) = \mathrm{E}(X+Y)^2 - \mathrm{E}^2(X+Y)= \mathrm{E}(X^2+Y^2+2XY) - (\mathrm{E}^2X+\mathrm{E}^2Y+2\mathrm{E}X\mathrm{E}Y)=$

$\mathrm{E}X^2+\mathrm{E}Y^2+2\mathrm{E}XY-\mathrm{E}^2X-\mathrm{E}^2Y-2\mathrm{E}X\mathrm{E}Y=$

$\mathrm{E}X^2-\mathrm{E}^2X+\mathrm{E}Y^2-\mathrm{E}^2Y+2(\mathrm{E}XY-\mathrm{E}X\mathrm{E}Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$.

Τέλος απόδειξης

Πόρισμα 1.2 Αν X,Y ανεξάρτητες τότε $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$, και $\mathrm{Var}(\sum X\mathrm{i}) = \sum \mathrm{Var}(X\mathrm{i})$ (Η γενίκευση με επαγωγή)

1.3 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ ΟΜΟΙΟΜΟΡΦΗ ΔΙΑΚΡΙΤΗ ΚΑΤΑΝΟΜΗ

Αν *X* μία τυχαία μεταβλητή η οποία παίρνει τις τιμές α,α+1,α+2,…,β για α, β∈ℤ με σταθερή πιθανότητα $\frac{1}{\beta-\alpha+1}$, τότε λέμε ότι η *X* ακολουθεί την **διακριτή ομοιόμορφη κατανομή** με παράμετρο β-α+1 και γράφουμε $\boldsymbol{X\sim U(\alpha, \beta)}$. Η συνάρτηση πιθανότητας της *X* είναι η :

$P(x) = \frac{1}{\beta-\alpha+1}$ x=α, α+1,α+2,…….,β.

Προφανώς p(x)≥ 0 ∀x και $\sum p(x) = \frac{\beta-\alpha+1}{\beta-\alpha+1} = 1$.

Άρα μ= E*X*=$\sum_{x=\alpha}^{\beta} x\, p(x) = \sum_{x=\alpha}^{\beta} \frac{x}{\beta-\alpha+1} = \frac{1}{\beta-\alpha+1} \sum_{x=\alpha}^{\beta} x$.

<u>Πρόταση 1.6</u> Από τύπο αριθμητικής προόδου έχουμε $S_n = \frac{n}{2}(a_1 + a_n)$ για το άθροισμα των n πρώτων όρων μιας αριθμητικής προόδου με a_1 τον 1^o όρο και a_n τον τελευταίο.

Απόδειξη

Αν ω η διαφορά της προόδου τότε

$S_{n=}\ a_1 + (a_1+w)+(a_1+2w)+…..+(a_1+(n-1)w)$ και

$S_{n=}\ a_n + (a_n-w)+(a_n-2w)+…..+(a_n-(n-1)w)$.

Προσθέτοντας τις δύο τελευταίες σχέσεις κατά μέλη έχουμε $2S_n=n(a_1+a_n)$,οπότε

$S_n = \frac{n}{2}(a_1 + a_n)$.

Τέλος απόδειξης.

Συνεπώς E*X*= $\frac{1}{\beta-\alpha+1}\ \frac{\beta-\alpha+1}{2}\ (\alpha+\beta)= \frac{\alpha+\beta}{2}$.

Για την διασπορά ή διακύμανση έχουμε Var*X*= $EX^2 - (EX)^2 = EX^2 - \frac{(a+b)^2}{4}$.

$EX^2 = \sum_{x=\alpha}^{\beta} x^2\, p(x) = \sum_{x=\alpha}^{\beta} \frac{x^2}{\beta-\alpha+1} = \frac{1}{\beta-\alpha+1} \sum_{x=\alpha}^{\beta} x^2.$

<u>Πρόταση 1.7</u> $\sum_{\iota=1}^{K} \iota^2 = \frac{K(K+1)(2K+1)}{6}$.

Απόδειξη

Με επαγωγή για K=1 έχουμε $1^2 = \frac{1\times 2\times 3}{6}$ που ισχύει.

Έστω ότι ισχύει για K. Αρκεί να δείξω ότι ισχύει για K+1.

Είναι $\sum_{\iota=1}^{K+1} \iota^2 = \sum_{\iota=1}^{K} \iota^2 + (K+1)^2 = \frac{K(K+1)(2K+1)}{6} + (K+1)^2 = \frac{K(K+1)(2K+1)+6(K+1)^2}{6}$

$=\frac{(K+1)[K(2K+1)+6(K+1)]}{6}=\frac{(K+1)(2K^2+K+6K+6)}{6}= \frac{(K+1)(2K^2+7K+6)}{6} =\frac{(K+1)(K+2)(2K+3)}{6}=$

$\frac{(K+1)(K+2)[2(K+1)+1]}{6}$ οπότε η πρόταση ισχύει για K+1.

Τέλος απόδειξης.

Συνεπώς $\sum_{x=a}^{\beta} x^2 = \sum_{x=1}^{\beta} x^2 - \sum_{x=1}^{a-1} x^2 = \frac{\beta(\beta+1)(2\beta+1)}{6} - \frac{(\alpha-1)\alpha(2\alpha-1)}{6} =$

$\frac{(\beta^2+\beta)(2\beta+1)-[(\alpha^2-\alpha)(2\alpha-1)]}{6} =\frac{2\beta^3+\beta^2+2\beta^2+\beta-(2\alpha^3-\alpha^2-2\alpha^2+\alpha)}{6}$

$=\frac{2\beta^3+\beta^2+2\beta^2+\beta-2\alpha^3+\alpha^2+2\alpha^2-\alpha}{6}$.

Άρα $EX^2 = \frac{2\beta^3+\beta^2+2\beta^2+\beta-2\alpha^3+3\alpha^2-\alpha}{6(\beta-\alpha+1)}$.

Οπότε $\mathrm{Var}X = EX^2 - (EX)^2 = \frac{2\beta^3+\beta^2+2\beta^2+\beta-2\alpha^3+3\alpha^2-\alpha}{6(\beta-\alpha+1)} - \frac{(a+b)^2}{4} =$

$\frac{4\beta^3+2\beta^2+4\beta^2+2\beta-4\alpha^3+6\alpha^2-2\alpha}{12(\beta-\alpha+1)} -3\frac{(a^2+2ab+b^2)(b-a+1)}{12(b-a+1)}$

$=\frac{4\beta^3+2\beta^2+4\beta^2+2\beta-4\alpha^3+6\alpha^2-2\alpha-3(a^2b-a^3+a^2+2ab^2-2a^2b+2ab+b^3-b^2a+b^2)}{12(\beta-\alpha+1)}=$

$$=\frac{\beta^3-a^3+3\beta^2+3a^2+2\beta-2a-6\alpha\beta+3\alpha^2\beta-3\alpha\beta^2}{12(\beta-\alpha+1)}=$$

$$\frac{\beta^3-\alpha^3+2\beta^2+\beta^2+2\alpha^2+\alpha^2+2\beta-2\alpha-2\alpha\beta-2\alpha\beta-2\alpha\beta+2\alpha^2\beta+\alpha^2\beta-2\alpha\beta^2-\alpha\beta^2}{12(\beta-\alpha+1)}=$$

$$\frac{\beta^3+2\beta^2-2\alpha\beta+\alpha^2\beta-2\alpha\beta^2}{12(\beta-\alpha+1)}+\frac{-\alpha^3+2\alpha^2-2\alpha\beta+2\alpha^2\beta-\alpha\beta^2}{12(\beta-\alpha+1)}+\frac{\beta^2+\alpha^2+2\beta-2\alpha-2\alpha\beta}{12(\beta-\alpha+1)}=$$

$$\frac{\beta(\beta^2+2\beta-2\alpha+\alpha^2-2\alpha\beta)}{12(\beta-\alpha+1)}-\frac{\alpha(\alpha^2-2\alpha+2\beta-2\alpha\beta+\beta^2)}{12(\beta-\alpha+1)}+\frac{\beta^2+\alpha^2+2\beta-2\alpha-2\alpha\beta}{12(\beta-\alpha+1)}=$$

$$\frac{\beta^2+\alpha^2+2\beta-2\alpha-2\alpha\beta}{12(\beta-\alpha+1)}\quad(\beta\text{-}\alpha+1)=\frac{\beta^2+\alpha^2+2\beta-2\alpha-2\alpha\beta}{12}=\frac{(\beta-\alpha)^2+2(\beta-\alpha)}{12}=\frac{(\beta-\alpha)(\beta-\alpha+2)}{12}=$$

$$\frac{(\beta-\alpha+1-1)(\beta-\alpha+1+1)}{12}=\frac{(\beta-\alpha+1)^2-1}{12}.$$

Από τον τύπο φαίνεται ότι η διασπορά της ομοιόμορφης κατανομής εξαρτάται αποκλειστικά από το εύρος του διαστήματος [α, β] και όχι από τα άκρα του διαστήματος. Συνεπώς ο τύπος της διασποράς που αποδείχτηκε για α, β≥1 με α, β φυσικοί αριθμοί με $\alpha < \beta$ θα μπορεί να ισχύει και για α, β∈ℤ. Για παράδειγμα η διασπορά της ομοιόμορφης διακριτής κατανομής στο διάστημα [-5,-2] θα είναι ίδια με τη διασπορά στο διάστημα που θα προκύψει αν προσθέσουμε 6 μονάδες στα άκρα δηλαδή στο [1,4].

Άρα ο συντελεστής μεταβλητότητας θα είναι $CV=\frac{\sigma}{\mu}=\frac{\sqrt{\frac{(\beta-\alpha+1)^2-1}{12}}}{\frac{\alpha+\beta}{2}}=\sqrt{\frac{(\beta-\alpha+1)^2-1}{3(\alpha+\beta)^2}}$.

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **ομοιόμορφη διακριτή κατανομή στο [α ,β]** είναι:

$$\boldsymbol{\mu= EX=\frac{\alpha+\beta}{2}}$$

$$\boldsymbol{\sigma^2=VarX=\frac{(\beta-\alpha+1)^2-1}{12}}$$

$$\boldsymbol{CV=\frac{\sigma}{\mu}=\sqrt{\frac{(\beta-\alpha+1)^2-1}{3(\alpha+\beta)^2}}}$$

1.4 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ BERNOULLI ΚΑΤΑΝΟΜΗ

Μία τυχαία μεταβλητή X ακολουθεί **κατανομή Bernoulli** και γράφουμε $\boldsymbol{X \sim B(1,p)}$ όταν είναι δίτιμη δηλαδή παίρνει μόνο δύο τιμές. Αυτές είναι συνήθως το 0 και το 1 και τις οποίες ονομάζουμε επιτυχία και αποτυχία, με πιθανότητα που δίνεται από τον τύπο:

$P(X=x)= p^x (1-p)^{1-x}$, x=0,1 και 0< p <1. Προφανώς $p(x) \geq 0 \ \forall x$, αφού p(0)=1-p και p(1)= p.

Επίσης $\sum p(x) = p(0) + p(1) = 1-p + p =1$.Άρα $EX= \sum_{x=0}^{1} xp(x) = 0p(0) + 1p(1)=p$.

$VarX=EX^2 – (EX)^2 = \sum_{x=0}^{1} x^2 p(x) - p^2 = 0p(0) + 1^2p(1) – p^2 = p-p^2 =p(1-p)$.

Συνεπώς $CV=\frac{\sigma}{\mu} = \frac{\sqrt{p(1-p)}}{p} \sqrt{\frac{1-p}{p}} = \sqrt{\frac{1}{p} - 1}$.

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **διακριτή κατανομή Bernoulli B(1,p)** είναι:

$\boldsymbol{\mu = EX=p}$

$\boldsymbol{\sigma^2=VarX= p(1-p)}$

$\boldsymbol{CV=\frac{\sigma}{\mu}=\sqrt{\frac{1}{p} - 1}}$

1.5 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Η **διωνυμική κατανομή B(n, p)** είναι μια διαδικασία n ανεξάρτητων δοκιμών Bernoulli με την ίδια πιθανότητα επιτυχίας p σε κάθε δοκιμή. Η τυχαία μεταβλητή X που ακολουθεί διωνυμική κατανομή($X \sim B(n,p)$), εκφράζει το πλήθος των επιτυχιών στις n δοκιμές(επαναλήψεις) Bernoulli και παίρνει τιμές x=0,1,……,n. Η συνάρτηση πιθανότητας δίνεται από τον τύπο: $P(X=x)= \binom{n}{x} p^x (1-p)^{n-x}$, x=0,1,……,n, 0< p <1 .

Προφανώς $p(x) \geq 0 \ \forall x$ και $\sum p(x)=\sum \binom{n}{x} p^x(1-p)^{n-x} =(p+1-p)^n =1$

(διώνυμο του Νεύτωνα).

Για τη μέση τιμή έχουμε μ= EX= $\sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$ =

$np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}$ =

$np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} = np\sum_{i=0}^{n} \binom{n-1}{i} p^i (1-p)^{n-1-i}$ =

np [p+(1-p)]=np .

Παραπάνω χρησιμοποιήθηκε το διώνυμο του Νεύτωνα.

<u>Πρόταση 1.8</u> Διώνυμο του Νεύτωνα $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^{n-k} b^k$ $\forall$n φυσικό αριθμό και a,b$\in\mathbb{R}$.

Απόδειξη

Έστω P(n) η πρόταση $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^{n-k} b^k$.

Θα αποδείξουμε ότι ισχύει με τη μέθοδο της επαγωγής.

Η P(0) ισχύει καθώς $(a+b)^0= 1=\binom{0}{0}a^0 b^0$

Έστω ότι ισχύει η P(n) δηλαδή ότι $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^{n-k} b^k$.

Αρκεί να δείξω ότι ισχύει η P(n+1) δηλαδή ότι $(a+b)^{n+1} = \sum_{k=0}^{n+1}\binom{n+1}{k}a^{n+1-k} b^k$.

Είναι $(a+b)^{n+1} = (a+b)^n (a+b)= \sum_{k=0}^{n}\binom{n}{k}a^{n-k} b^k$ (a+b)=

$[\binom{n}{0}a^n + \binom{n}{1} a^{n-1} b + \binom{n}{2} a^{n-2} b^2 +\ldots\ldots\ldots\ldots+\binom{n}{\lambda} a^{n-\lambda} b^{\lambda} +\ldots\ldots\ldots+ \binom{n}{n-2} a^2 b^{n-2} + \binom{n}{n-1} a b^{n-1} +\binom{n}{n} b^n]$ (a+b)=

$\binom{n}{0}a^{n+1} +\binom{n}{0}a^n b + \binom{n}{1} a^n b + \binom{n}{1} a^{n-1} b^2 + \binom{n}{2} a^{n-1} b^2 + \binom{n}{2} a^{n-2} b^3 +\ldots\ldots\ldots\ldots+\binom{n}{\lambda} a^{n-\lambda+1} b^{\lambda} + \binom{n}{\lambda} a^{n-\lambda} b^{\lambda+1} + \binom{n}{\lambda+1} a^{n-\lambda} b^{\lambda+1} + \binom{n}{\lambda+1} a^{n-\lambda-1} b^{\lambda+2} +\ldots\ldots\ldots+ \binom{n}{n-2} a^2 b^{n-1} +\binom{n}{n-1} a^2 b^{n-1} + \binom{n}{n-1} a b^n + \binom{n}{n} ab^n +\binom{n}{n} b^{n+1}$ =

$\binom{n}{0}a^{n+1} + [\binom{n}{0} + \binom{n}{1}]\ a^n\ b + [\binom{n}{1} + \binom{n}{2}]\ a^{n-1}\ b^2 + \ldots\ldots\ldots + [\binom{n}{\lambda} + \binom{n}{\lambda+1}]\ a^{n-\lambda}\ b^{\lambda+1}$
$+\ldots\ldots.. + [\binom{n}{n-2} + \binom{n}{n-1}]\ a^2\ b^{n-1} + [\binom{n}{n-1} + \binom{n}{n}]\ a\ b^n + \binom{n}{n}\ b^{n+1}$

(Σχέση 1.2).

Όμως $\binom{n}{0} = \binom{n+1}{0}$ **(Σχέση 1.3)** γιατί $\binom{n}{0} = \frac{n!}{0!n!} = 1$ και $\binom{n+1}{0} = \frac{(n+1)!}{0!(n+1)!} = 1$.

Επίσης $\binom{n}{n} = \binom{n+1}{n+1}$ **(Σχέση 1.4)** γιατί $\binom{n}{n} = \frac{n!}{n!0!} = 1$ και $\binom{n+1}{n+1} = \frac{(n+1)!}{(n+1)!0!} = 1$.

Επίσης έχουμε $\binom{n}{\lambda-1} + \binom{n}{\lambda} = \binom{n+1}{\lambda}$ **(Σχέση 1.5)** καθώς

$$\binom{n}{\lambda-1} + \binom{n}{\lambda} = \frac{n!}{(\lambda-1)!(n-\lambda+1)!} + \frac{n!}{\lambda!(n-\lambda)!} = \frac{n!}{(\lambda-1)!(n-\lambda)!(n-\lambda+1)} + \frac{n!}{(\lambda-1)!\,\lambda\,(n-\lambda)!} =$$

$$n!\left(\frac{1}{(n-\lambda+1)(\lambda-1)!(n-\lambda)!} + \frac{1}{\lambda\,(\lambda-1)!\,(n-\lambda)!}\right) = n!\,\frac{\lambda+n-\lambda+1}{\lambda(n-\lambda+1)(\lambda-1)!(n-\lambda)!} = n!\,\frac{n+1}{\lambda!(n-\lambda+1)!} =$$

$$\frac{(n+1)!}{\lambda!(n-\lambda+1)!} = \binom{n+1}{\lambda}.$$

Συνεπώς από σχέσεις (1.2), (1.3), 1.4), (1.5) έχουμε

$(a+b)^{n+1} = \binom{n+1}{0}a^{n+1} + \binom{n+1}{1}\ a^n\ b + \binom{n+1}{2}\ a^{n-1}\ b^2 + \ldots\ldots\ldots + \binom{n+1}{\lambda+1}\ a^{n-\lambda}\ b^{\lambda+1} + \ldots\ldots.. +$
$\binom{n+1}{n-1}\ a^2\ b^{n-1} + \binom{n+1}{n}\ a\ b^n + \binom{n+1}{n+1}\ b^{n+1} = \sum_{k=0}^{n+1}\binom{n+1}{k}a^{n+1-k}\ b^k$, δηλαδή ισχύει η P(n+1).

Τέλος απόδειξης.

Για τη παραγοντική ροπή δεύτερης τάξης έχουμε EX(X-1)=

$$\sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x\,(1-p)^{n-x} = \sum_{x=2}^{n} x(x-1)\,\frac{n!}{x!(n-x)!}\,p^x\,(1-p)^{n-x} =$$

$$n(n-1)p^2\ \sum_{x=2}^{n}\frac{(n-2)!}{(x-2)!(n-x)!}\,p^{x-2}\,(1-p)^{n-x} = n(n-1)p^2\ \sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}\,(1-p)^{n-x} =$$

$$n(n-1)p^2\ \sum_{i=0}^{n}\binom{n-2}{i}p^i\,(1-p)^{n-2-i} = n(n-1)p^2\ (p+1-p)^{n-2} = n(n-1)p^2\,.$$

Όμως EX^2 = EX(X-1) + EX = n(n-1)p^2 + np =$n^2p^2 - np^2 + np = n^2p^2 + np(1-p)$.

Άρα Var$X$= EX^2 –(EX)^2 = $n^2p^2 + np(1-p) - n^2p^2 = np(1-p)$.

Συνεπώς CV=$\frac{\sigma}{\mu} = \frac{\sqrt{np(1-p)}}{np} = \sqrt{\frac{1-p}{np}} = \sqrt{\frac{1}{np} - \frac{1}{n}} = \sqrt{\frac{1}{n}(\frac{1}{p} - 1)}$.

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **διωνυμική κατανομή B(n,p)** είναι:

μ =E*X*=np

σ^2=Var*X*= np(1-p)

CV=$\frac{\sigma}{\mu}$=$\sqrt{\frac{1}{n}(\frac{1}{p} - 1)}$

1.6 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ ΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Η τυχαία μεταβλητή *X* με συνάρτηση πιθανότητας P(*X*=x)=p(x)=p(1-p)^x , x=0,1,2,…. ,όπου 0< p <1 λέμε ότι ακολουθεί τη **γεωμετρική κατανομή.**

Προφανώς p(x) ≥ 0 ∀x και $\sum p(x) = \sum p(1-p)^x = p\frac{1}{1-(1-p)} = \frac{p}{p} = 1$.

<u>Πρόταση 1.9</u> Η πιθανογεννήτρια της γεωμετρικής κατανομής με παράμετρο p είναι :

$\Pi_x(t) = \frac{p}{1-(1-p)t}$,$|t| < 1$.

Απόδειξη

$\Pi_x(t) = Et^x = \sum_{x=0}^{\infty} t^x\, p(1-p)^x = p\sum_{x=0}^{\infty}[(1-p)t]^x = \frac{p}{1-(1-p)t}$, αν |(1-p)t|<1.

Τέλος απόδειξης

Παραπάνω χρησιμοποιήσαμε τον τύπο της γεωμετρικής σειράς :

$\sum_{x=0}^{\infty} \alpha^x = \frac{1}{1-a}$ για |α|< 1.

Βέβαια αυτός ο τύπος προκύπτει από τον τύπο $\sum_{x=0}^{\nu} \alpha^x = \frac{1-\alpha^\nu}{1-a}$ όπου $\lim_{\nu\to\infty} \alpha^\nu = 0$ για |α|< 1.

<u>Πρόταση 1.10</u> Το άθροισμα S_n ,των n πρώτων όρων γεωμετρικής προόδου είναι

$S_n= \frac{\alpha 1(\lambda^{\nu}-1)}{\lambda-1}$, όπου λ ο λόγος της προόδου και α_1 ο $1^{ος}$ όρος.

Απόδειξη

$S_n = \alpha_1 + \alpha_1\lambda+\alpha_1\lambda^2 + ...+ \alpha_1\lambda^{\nu-1}$.

$\lambda S_n=\alpha_1\lambda+ \alpha_1\lambda^2+\alpha_1\lambda^3 + ...+ \alpha_1\lambda^{\nu}$.

Αφαιρούμε από την τελευταία σχέση την προτελευταία και έχουμε:

$\lambda S_n - S_n = \alpha_1\lambda^{\nu} - \alpha_1$. Συνεπώς $S_n= \frac{\alpha 1(\lambda^{\nu}-1)}{\lambda-1}$.

Τέλος απόδειξης

Από τη σχέση E*X*=Π΄(1), θα υπολογίσουμε τη μέση τιμή.

Είναι $\Pi'(t)=-\frac{p}{[1-(1-p)t]^2}$ (p - 1) , οπότε μ= E*X*=$\Pi'(1) = \frac{p(1-p)}{p^2}=\frac{1-p}{p}$.

Επίσης από τη σχέση σ^2= Var*X*=Π΄΄(1) + Π΄(1) –(Π΄(1))^2 θα υπολογίσουμε την διασπορά.

Είναι $\Pi''(t) = \frac{(p-1)p}{[1-(1-p)t]^4}$ 2[1-(1-p)t] (p-1),οπότε $\Pi''(1) =\frac{(p-1)p\ 2p(p-1)}{p^4}=\frac{2(p-1)^2}{p^2}$.

Άρα $\sigma^2 = \frac{2(p-1)^2}{p^2} + \frac{1-p}{p} - \frac{(1-p)^2}{p^2} = \frac{(p-1)^2}{p^2} + \frac{p(1-p)}{p^2} = \frac{1-p}{p^2}$.

Συνεπώς CV= $\frac{\sigma}{\mu}=\frac{\sqrt{\frac{1-p}{p^2}}}{\frac{1-p}{p}} = \sqrt{\frac{1}{1-p}} = \frac{1}{\sqrt{1-p}}$.

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **γεωμετρική κατανομή** είναι:

$\boldsymbol{\mu = EX=\frac{1-p}{p}}$

$\boldsymbol{\sigma^2=VarX= \frac{1-p}{p^2}}$

$$\mathbf{CV=\frac{\sigma}{\mu}=\frac{1}{\sqrt{1-p}}}$$

1.7 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ POISSON ΚΑΤΑΝΟΜΗ

Η τυχαία μεταβλητή X που ακολουθεί την **κατανομή Poisson** με παράμετρο λ, ($X \sim P(\lambda)$) έχει συνάρτηση πιθανότητας $P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x=0,1,2,\ldots..$, $\lambda>0$ (παράμετρος) .

Προφανώς $p(x) \geq 0\ \forall x$ και $\sum p(x) = \sum e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum\frac{\lambda^x}{x!} = e^{-\lambda}\ e^{\lambda} = 1$.

Η κατανομή Poisson είναι η κατανομή των σπάνιων γεγονότων και χρησιμοποιείται όταν θέλουμε να μετρήσουμε τον αριθμό των "συμβάντων" στην μονάδα μέτρησης. Τα συμβάντα μπορεί να είναι τυπογραφικά λάθη ανά σελίδα, τηλεφωνικές κλήσεις στη μονάδα του χρόνου, αυτοκίνητα που περνούν από ένα σημείο στη μονάδα του χρόνου κ.α.

<u>Πρόταση 1.11</u> Αν η τυχαία μεταβλητή X ακολουθεί την κατανομή Poisson τότε η πιθανογεννήτρια της θα είναι η $\Pi_x(z) = e^{\lambda(z-1)}$.

Απόδειξη

$\Pi_x(z) = E(z^x) = \sum_{x=0}^{\infty} p(x)z^x = \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x z^x}{x!} = e^{-\lambda}\ e^{\lambda z} = e^{\lambda(z-1)}$.

Τέλος απόδειξης

<u>Πρόταση 1.12</u> Ισχύει η σχέση $\sum_{x=0}^{\infty}\frac{\alpha^x}{x!} = e^{\alpha}$.

Απόδειξη

Από ανάπτυγμα Taylor (θεωρείται γνωστό) ισχύει ότι αν $f(x)=e^x$,τότε

$$f(x)= f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \ldots\ldots\ldots + \frac{f^{(n)}(0)}{n!}x^n + Rn(x) .$$

Όμως η $f(x)=e^x$ έχει παραγώγους κάθε τάξης στο διάστημα $I=[0,x]$, με $0 \in I$ και οι παράγωγοι $f^{(n)}(x)$ είναι ομοιόμορφα φραγμένοι στο I ,δηλαδή ισχύει ότι

$\exists M > 0$,$\forall$ n $\in$ N, $\forall$ x $\in$ I : $|f^{(n)}(x)| < M$. **(Σχέση 1.6) .**

Επιλέγουμε ως Rn(x) το υπόλοιπο του Lagrange

$R_n^L(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\ x^{n+1}$, $\xi \in (0,x)$,$\forall$ n $\in$ N, $\forall$ x $\in$ I.

Από σχέση 1.6 έχουμε ότι $\forall$ n $\in$ N, $\forall$ x $\in$ I : $| R_n^L(x)| \leq \frac{M}{(n+1)!} |x^{n+1}|$.

Οπότε $\lim_{n\to\infty} \frac{M}{(n+1)!}\ |x^{n+1}| = 0$ καθώς η σειρά $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ συγκλίνει .

Οπότε $\lim_{n\to\infty}$ RnL(x) =0 , $\forall$ x $\in$ I.(από κριτήριο παρεμβολής που θεωρείται γνωστό).

Συνεπώς f(x)= $\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}\ x^n$, δηλαδή

$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ ή $e^{\alpha} = \sum_{x=0}^{\infty} \frac{\alpha^x}{x!}$.

Τέλος απόδειξης

<u>Πρόταση 1.13</u> Στην προηγούμενη απόδειξη χρησιμοποιήσαμε ότι η σειρά $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ συγκλίνει. Αυτό προκύπτει από κριτήριο D Alembert.

Απόδειξη

Σύμφωνα με το κριτήριο D Alembert(το θεωρούμε γνωστό) αν έχουμε τη σειρά θετικών όρων $\sum_{n=1}^{\infty} an$, $\alpha_n > 0$, $\forall$ n $\in$ N τότε αν

$\lim_{n\to\infty} \frac{an+1}{an} < 1$ η σειρά $\sum_{n=1}^{\infty} an$ συγκλίνει.

Στην συγκεκριμένη περίπτωση $\lim_{n\to\infty} \frac{an+1}{an} = \lim_{n\to\infty} \frac{\frac{x^{n+1}}{(n+1)!}}{\frac{x^n}{n!}} = \lim_{n\to\infty} \frac{x}{n+1} = 0 < 1$.

Επίσης θεωρούμε γνωστό ότι αν η σειρά $\sum_{n=1}^{\infty} an$ συγκλίνει τότε $\alpha_n \to 0$ και ολοκληρώνουμε την απόδειξη.

Τέλος απόδειξης

Συνεπώς από τη σχέση μ= EX=Π΄(1) θα υπολογίσουμε τη μέση τιμή.

Είναι $\Pi_x(z) = e^{\lambda(z-1)}$, οπότε $\Pi'_x(z) = \lambda e^{\lambda(z-1)}$, άρα Π΄(1)=λ.

Επίσης από τη σχέση σ^2= Var$X$=Π΄΄(1) + Π΄(1) – (Π΄(1))^2 θα υπολογίσουμε την διασπορά.

Είναι $\Pi_x''(z) = \lambda^2 e^{\lambda(z-1)}$, οπότε Π΄΄(1)= λ^2 .

Τελικά σ^2= Var$X$= λ^2 + λ - λ^2 =λ και CV= $\frac{\sigma}{\mu} = \frac{\sqrt{\lambda}}{\lambda} = \frac{1}{\sqrt{\lambda}}$.

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **κατανομή Poisson** είναι:

μ =EX= λ

σ^2=Var$X$= λ

CV= $\frac{\sigma}{\mu} = \frac{1}{\sqrt{\lambda}}$

1.8 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ ΑΡΝΗΤΙΚΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Θα λέμε ότι η τυχαία μεταβλητή X έχει την **αρνητική διωνυμική κατανομή** αν έχει συνάρτηση πιθανότητας P(X=x) = p(x) =$\binom{\nu+x-1}{x}$ p^ν $(1-p)^x$, x=0,1,…..

όπου $\nu > 0$ και p παράμετρος με 0 < p <1. Προφανώς p(x)≥ 0 και $\sum$p(x) =1 (θα το δείξουμε παρακάτω).

Πριν προχωρήσουμε στην εύρεση των αριθμητικών μέτρων που μας ενδιαφέρουν θα αποδείξουμε κάποιες σχέσεις που θα μας βοηθήσουν για την απόδειξη του τύπου της Πιθανογεννήτριας.

<u>Πρόταση 1.14</u> Ισχύει ότι $\binom{-\alpha}{\kappa}$ =$(-1)^\kappa$ $\binom{\alpha+\kappa-1}{\kappa}$ με α>0 και κ φυσικό.

Απόδειξη

$\binom{-\alpha}{\kappa} = \frac{-\alpha(-\alpha-1)\ldots\ldots\ldots(-\alpha-\kappa-1)}{\kappa!} = \frac{(-1)^\kappa \alpha(\alpha+1)\ldots\ldots(\alpha+\kappa-1)}{\kappa!} =$

$(-1)^{\kappa}\frac{(\alpha+\kappa-1)(\alpha+\kappa-2)\ldots\ldots\ldots\alpha}{\kappa!} = (-1)^{\kappa}\binom{\alpha+\kappa-1}{\kappa}$.

Τέλος απόδειξης

<u>Πρόταση 1.15</u> ισχύει ότι $\binom{\alpha+\kappa}{\kappa} = (-1)^{\kappa}\binom{-\alpha-1}{\kappa}$.

Απόδειξη

Θέτοντας β=α+1 έχουμε

$\binom{\alpha+\kappa}{\kappa} = \binom{\beta+\kappa-1}{\kappa} = (-1)^{\kappa}\binom{-\beta}{\kappa}$ λόγω πρότασης 1.14.

Έτσι $\binom{\alpha+\kappa}{\kappa} = (-1)^{\kappa}\binom{-\alpha-1}{\kappa}$.

Τέλος απόδειξης

<u>Πρόταση 1.16</u> Ισχύει $\sum_{x=0}^{\infty}\binom{\nu+x-1}{x}\alpha^{\nu}\beta^{x} = \alpha^{\nu}(1-\beta)^{-\nu}$.

Απόδειξη

$\sum_{x=0}^{\infty}\binom{\nu+x-1}{x}\alpha^{\nu}\beta^{x} = \alpha^{\nu}\sum_{x=0}^{\infty}\binom{\nu+x-1}{x}\beta^{x} = \alpha^{\nu}\sum_{x=0}^{\infty}(-1)^{x}\binom{-\nu}{x}\beta^{x}$ λόγω πρότασης(1.14).

Άρα $\sum_{x=0}^{\infty}\binom{\nu+x-1}{x}\alpha^{\nu}\beta^{x} = \alpha^{\nu}\sum_{x=0}^{\infty}\binom{-\nu}{x}(-\beta)^{x} = \alpha^{\nu}(1-\beta)^{-\nu}$ από διώνυμο του Νεύτωνα.

Τέλος απόδειξης

Άρα $\sum p(x) = \sum\binom{\nu+x-1}{x}p^{\nu}(1-p)^{x} = p^{\nu}(1-1+p)^{-\nu} = 1$.

<u>Πρόταση 1.17</u> Η Πιθανογεννήτρια της τυχαίας μεταβλητής *X* που ακολουθεί την αρνητική διωνυμική κατανομή είναι $\Pi_x(t) = =\left(\frac{p}{1-(1-p)t}\right)^{\nu}$.

Απόδειξη

Είναι $\Pi_x(t) = Et^{x} = \sum_{x=0}^{\infty} t^{x}\binom{\nu+x-1}{x}p^{\nu}(1-p)^{x} = \sum_{x=0}^{\infty}\binom{\nu+x-1}{x}p^{\nu}[(1-p)t]^{x} =$

$p^{\nu}\,(1-(1-p)t)^{-\nu} = \left(\frac{p}{1-(1-p)t}\right)^{\nu}$.

Τέλος απόδειξης

Θα βρούμε τη μέση τιμή από τον τύπο μ= EX=Π΄(1).

Είναι $\Pi_x(t) = \left(\frac{p}{1-(1-p)t}\right)^{\nu}$, άρα $\Pi'_x(t) = \nu \left(\frac{p}{1-(1-p)t}\right)^{\nu-1} p \left(\frac{-1}{(1-(1-p)t)^2}\right)(p-1) =$

$\nu p\,(1-p)\frac{p^{\nu-1}}{(1-(1-p)t)^{\nu+1}} = \nu(1-p)\frac{p^{\nu}}{(1-(1-p)t)^{\nu+1}}$. Συνεπώς μ= EX=Π΄(1)= $\frac{\nu(1-p)}{p}$.

Επίσης από τη σχέση σ^2= Var$X$=Π΄΄(1) + Π΄(1) – (Π΄(1))^2 θα υπολογίσουμε την διασπορά.

Είναι $\Pi_x''(t) = \nu(1-p)\frac{p^{\nu}}{-(1-(1-p)t)^{2\nu+2}}\ (\nu+1)\ (1-(1-p)t)^{\nu}\ (p-1) = \frac{\nu(1-p)^2 p^{\nu}(\nu+1)}{(1-(1-p)t)^{\nu+2}}$.

Άρα $\Pi''(1) = \frac{\nu(1-p)^2(\nu+1)}{p^2}$.

Συνεπώς σ^2= Var$X$= $\frac{\nu(1-p)^2(\nu+1)}{p^2} + \frac{\nu(1-p)}{p} - \frac{\nu^2(1-p)^2}{p^2} = \frac{(1-p)^2(\nu^2+\nu-\nu^2)}{p^2} + \frac{\nu p(1-p)}{p^2} =$

$\frac{\nu+\nu p^2-2p\nu+\nu p-\nu p^2}{p^2} = \frac{\nu-\nu p}{p^2} = \frac{\nu(1-p)}{p^2}$

Άρα CV= $\frac{\sigma}{\mu} = \frac{\sqrt{\frac{\nu(1-p)}{p^2}}}{\frac{\nu(1-p)}{p}} = \frac{1}{\sqrt{\nu(1-p)}}$.

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **αρνητική διωνυμική κατανομή** είναι:

μ =EX= $\frac{\nu(1-p)}{p}$.

σ^2=Var$X$= $\frac{\nu(1-p)}{p^2}$

CV= $\frac{\sigma}{\mu}$ =$\frac{1}{\sqrt{\nu(1-p)}}$.

1.9 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Η τυχαία μεταβλητή X που παίρνει τις τιμές 0,1,……,ν, θα λέμε ότι ακολουθεί την **υπεργεωμετρική κατανομή** αν έχει συνάρτηση πιθανότητας

$P(X=x) = p(x) = \frac{\binom{K}{x}\binom{N-K}{\nu-x}}{\binom{N}{\nu}}$,$0 \leq x \leq \nu$, όπου N, K, ν είναι ακέραιοι με $0 \leq K \leq N$,$1 \leq \nu \leq N$.

Για να είναι $p(x) \geq 0$ πρέπει να ισχύει $0 \leq x \leq K$ και $0 \leq \nu - x \leq N-K$ ή ισοδύναμα το x να ικανοποιεί την διπλή ανισότητα : $\max\{0, \nu - N+K\} \leq x \leq \min\{\nu, K\}$.

Ακόμη $\sum p(x) = 1$ (θα το δείξουμε παρακάτω).

Στην εξίσωση $(1+x)^{n+m} = (1+x)^n (1+x)^m$ αν χρησιμοποιήσουμε το διώνυμο του Νεύτωνα $(\alpha + \beta)^\nu = \sum_{\kappa=0}^{\nu}\binom{\nu}{\kappa} \alpha^\kappa \beta^{\nu-\kappa}$ έχουμε :

$$(1+x)^{n+m} = \sum_{r=0}^{n+m}\binom{n+m}{r} x^r = \sum_{k=0}^{n}\binom{n}{k} x^k \sum_{p=0}^{m}\binom{m}{p} x^p .$$

Αν εξισώσουμε τους συντελεστές του x^r στα δύο μέλη της ταυτότητας θα έχουμε

$$\binom{n+m}{r} = \sum_{k=0}^{r}\binom{n}{k}\binom{m}{p} \text{ με } k+p=r \text{ δηλαδή } \binom{n+m}{r} = \sum_{k=0}^{r}\binom{n}{k}\binom{m}{r-k}$$

(Σχέση 1.7) .

Συνεπώς $\sum p(x) = \sum_{x=0}^{\nu} \frac{\binom{K}{x}\binom{N-K}{\nu-x}}{\binom{N}{\nu}} = \frac{\binom{N}{\nu}}{\binom{N}{\nu}} = 1$ λόγω σχέσης (1.7).

Επίσης είναι $EX = \sum x\,p(x) = \sum_{x=0}^{\nu} x \frac{\binom{K}{x}\binom{N-K}{\nu-x}}{\binom{N}{\nu}} = \sum_{x=1}^{\nu}\binom{N}{\nu}^{-1} x \binom{N-K}{\nu-x} \frac{K}{x} \binom{K-1}{x-1} = K$

$$\binom{N}{\nu}^{-1} \sum_{x=1}^{\nu}\binom{N-K}{\nu-x} \binom{K-1}{x-1} = K \binom{N}{\nu}^{-1} \sum_{\iota=0}^{\nu}\binom{N-K}{\nu-\iota-1} \binom{K-1}{\iota} =$$

$$K \binom{N}{\nu}^{-1} \binom{N-1}{\nu-1} = K \frac{\nu!(N-\nu)!}{N!} \frac{(N-1)!}{(\nu-1)!(N-\nu)!} = \frac{\nu K}{N} .$$

Ομοίως $EX(X-1) = \sum x(x-1)\,p(x) = \sum_{x=2}^{\nu} x(x-1) \frac{\binom{K}{x}\binom{N-K}{\nu-x}}{\binom{N}{\nu}}$.

Όμως $\binom{K}{x} = \frac{K(K-1)}{x(x-1)} \binom{K-2}{x-2}$ **(Σχέση 1.8)** .

Από τις δύο τελευταίες σχέσεις έχουμε $E\mathit{X}(\mathit{X}\text{-}1) = \frac{\sum_{x=2}^{\nu} x(x-1)\frac{K(K-1)}{x(x-1)}\binom{K-2}{x-2}\binom{N-K}{\nu-x}}{\binom{N}{\nu}} =$

$\frac{K(K-1)\sum_{x=2}^{\nu}\binom{K-2}{x-2}\binom{N-K}{\nu-x}}{\binom{N}{\nu}} = \frac{K(K-1)\sum_{\iota=0}^{\nu}\binom{K-2}{\iota}\binom{N-K}{\nu-\iota-2}}{\binom{N}{\nu}}$.

Από την τελευταία εξίσωση και τις σχέσεις (1.7) και (1.8) έχουμε

$E\mathit{X}(\mathit{X}\text{-}1)= \frac{K(K-1)\binom{N-2}{\nu-2}}{\frac{N(N-1)}{\nu\ \nu-1}\binom{N-2}{\nu-2}} = \nu(\nu\text{-}1)\frac{K(K-1)}{N(N-1)}$.

Συνεπώς $\sigma^2=\mathrm{Var}\mathit{X}= E\mathit{X}(\mathit{X}\text{-}1) + E\mathit{X} - (E\mathit{X})^2 = \nu(\nu\text{-}1)\frac{K(K-1)}{N(N-1)} + \frac{\nu K}{N} - \frac{\nu^2K^2}{N^2} =$

$(\nu^2\text{-}\nu)\frac{NK^2-NK}{N^2(N-1)} + \frac{\nu KN-\nu^2K^2}{N^2} =$

$\frac{\nu^2NK^2-\nu^2NK-\nu NK^2+\nu NK}{N^2(N-1)} + \frac{\nu KN^2-\nu^2K^2N-\nu KN+\nu^2K^2}{N^2(N-1)} = \frac{-\nu^2NK-\nu NK^2+\nu KN^2+\nu^2K^2}{N^2(N-1)} =$

$\frac{\nu K(-\nu N-NK+N^2+\nu K)}{N^2(N-1)} = \frac{\nu K[N(N-\nu)-K(N-\nu)]}{N^2(N-1)} = \frac{\nu K(N-K)(N-\nu)}{N^2(N-1)} = \nu\frac{K}{N}\frac{N-K}{N}\frac{N-\nu}{N-1}$.

Άρα $CV= \frac{\sigma}{\mu} = \frac{\sqrt{\frac{\nu K(N-K)(N-\nu)}{N^2(N-1)}}}{\frac{\nu K}{N}} = \sqrt{\frac{(N-K)(N-\nu)}{\nu K(N-1)}}$.

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **υπεργεωμετρική κατανομή** είναι:

$\boldsymbol{\mu = E\mathit{X}= \frac{\nu K}{N}}$

$\boldsymbol{\sigma^2=\mathrm{Var}\mathit{X}= \nu\frac{K}{N}\frac{N-K}{N}\frac{N-\nu}{N-1}}$

$\boldsymbol{CV= \frac{\sigma}{\mu} = \sqrt{\frac{(N-K)(N-\nu)}{\nu K(N-1)}}}$

1.10 ΑΡΙΘΜΗΤΙΚΑ ΜΕΤΡΑ ΓΙΑ ΑΡΝΗΤΙΚΗ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Έστω μια κάλπη που περιέχει Κ άσπρα και Ν-Κ μαύρα σφαιρίδια. Εξάγουμε διαδοχικά ,ένα προς ένα χωρίς επανάθεση τα σφαιρίδια μέχρι να πάρουμε το ν-οστό άσπρο σφαιρίδιο, με ν≤ Κ. Αν η τυχαία μεταβλητή που μετρά το πλήθος των απαιτούμενων εξαγωγών, τότε η *X* ακολουθεί την αρνητική υπεργεωμετρική κατανομή. Η συνάρτηση πιθανότητας της Χ είναι:

$P(X= x) = p(x)= \frac{\binom{K}{\nu-1}\binom{N-K}{x-\nu}}{\binom{N}{x-1}} \frac{K-\nu+1}{N-x+1}$, x= ν, ν +1,…….,ν+Ν-Κ. Προφανώς $p(x) \geq 0$, $\forall x$.

Αν *Y* η τυχαία μεταβλητή που μετρά το πλήθος των μαύρων σφαιριδίων μέχρι να πάρουμε το ν-οστό άσπρο σφαιρίδιο τότε *Y*= *X*- ν. Η συνάρτηση πιθανότητας της τυχαίας μεταβλητής *Y* είναι:

$P(Y= y) = p(y)= \frac{\binom{K}{\nu-1}\binom{N-K}{y}}{\binom{N}{\nu+y-1}} \frac{K-\nu+1}{N-y-\nu+1}$, y=0,1,……,Ν-Κ.

Προφανώς E*Y*=E*X*- ν και Var*Y*=Var*X*.

Θα υπολογίσουμε τα παραπάνω αριθμητικά μέτρα για την τυχαία μεταβλητή *Y* και μέσω αυτών τα αριθμητικά μέτρα για την τυχαία μεταβλητή *X*.

Ισχύει $\sum_{y=0}^{N-K} p(y)=\sum_{y=0}^{N-K} \frac{\binom{K}{\nu-1}\binom{N-K}{y}}{\binom{N}{\nu+y-1}} \frac{K-\nu+1}{N-y-\nu+1} =$

$\sum_{y=0}^{N-K} \frac{K!(N-K)!(\nu+y-1)!(N-\nu-y+1)!(K-\nu+1)}{(\nu-1)!(K-\nu+1)!y!(N-K-y)!N!(N-y-\nu+1)}=$

$\frac{1}{\binom{N}{N-K}} \sum_{y=0}^{N-K} \binom{\nu+y-1}{y} \frac{(N-\nu-y)!}{(K-\nu)!(N-K-y)!}=$

$\frac{1}{\binom{N}{N-K}} \sum_{y=0}^{N-K} \binom{y+\nu-1}{y} \binom{N-\nu-y}{N-K-y}= \frac{\binom{N}{N-K}}{\binom{N}{N-K}} =1$, οπότε και $\sum_{x=\nu}^{\nu+N-K} p(x) =1$.

Παραπάνω χρησιμοποιήθηκε η παρακάτω πρόταση

<u>Πρόταση 1.18</u> Ισχύει $\sum_{j=0}^{k} \binom{j+m}{j} \binom{n-m-j}{k-j} = \binom{n+1}{k}$

Απόδειξη

$\sum_{j=0}^{k}\binom{j+m}{j}\binom{n-m-j}{k-j}=\sum_{j=0}^{k}(-1)^{j}\binom{-m-1}{j}(-1)^{k-j}\binom{k-n+m-1}{k-j}=$

$\sum_{j=0}^{k}(-1)^{k}\binom{-m-1}{j}\binom{k-n+m-1}{k-j}=(-1)^{k}\binom{k-n-2}{k}=(-1)^{k}\binom{k-(n+1)-1}{k}=\binom{n+1}{k}.$

Τέλος απόδειξης

Παραπάνω χρησιμοποιήθηκε η παρακάτω πρόταση

<u>Πρόταση 1.19</u> Ισχύει $\binom{n}{k}=(-1)^{k}\binom{k-n-1}{k}$

Απόδειξη

$(-1)^{k}\binom{k-n-1}{k}=(-1)^{k}\frac{(k-n-1)!}{k!(-n-1)!}=(-1)^{k}\frac{(n-1+1)(-n-1+2)\ldots\ldots(-n-1+k)}{k!}=\frac{n(n-1)\ldots\ldots(n+1-k)}{k!}=$

$\binom{n}{k}$

Τέλος απόδειξης

Συνεπώς $EY=\sum_{y=0}^{N-K}yp(y)=\frac{1}{\binom{N}{K}}\sum_{y=0}^{N-K}y\binom{y+\nu-1}{y}\binom{N-\nu-y}{N-K-y}=$

$\frac{\nu}{\binom{N}{K}}\left[\sum_{y=0}^{N-K}\frac{y+\nu}{\nu}\binom{y+\nu-1}{\nu-1}\binom{N-\nu-y}{N-K-y}\right]-\nu=$

$\frac{\nu}{\binom{N}{K}}\left[\sum_{y=0}^{N-K}\binom{y+\nu}{\nu}\binom{N-\nu-y}{N-K-y}\right]-\nu=\frac{\nu}{\binom{N}{K}}\left[\sum_{y=0}^{N-K}\binom{y+\nu}{y}\binom{N-\nu-y}{N-K-y}\right]-\nu=$

$\frac{\nu}{\binom{N}{K}}\binom{N+1}{N-K}-\nu=\frac{\nu\,K!(N-K)!(N+1)!}{N!(N-K)!(K+1)!}-\nu=\nu\frac{N+1}{K+1}-\nu\frac{K+1}{K+1}=\nu\frac{N-K}{K+1}$

Συνεπώς από $EY=EX-\nu$ έχουμε ότι $EX=\nu\frac{N-K}{K+1}+\nu\frac{K+1}{K+1}=\nu\frac{N+1}{K+1}$

Επίσης έχουμε

$EY^{2}=\sum_{y=0}^{N-K}y^{2}\,p(y)=\sum_{y=0}^{N-K}(y+\nu)(y+\nu+1)p(y)-(2\nu+1)EY-\nu^{2}-\nu=$

$\frac{1}{\binom{N}{K}}\sum_{y=0}^{N-K}(y+\nu)(y+\nu+1)\binom{y+\nu-1}{y}\binom{N-\nu-y}{N-K-y}-(2\nu+1)EY-\nu^{2}-\nu=$

$= \frac{\nu(\nu+1)}{\binom{N}{K}} \sum_{y=0}^{N-K} \binom{y+\nu+1}{y} \binom{N-\nu-y}{N-K-y}$ - $(2\nu+1)EY$ - ν^2- ν =

$= \frac{\nu(\nu+1)}{\binom{N}{K}} \binom{N+2}{N-K}$ - $(2\nu+1)\,\nu\,\frac{N-K}{K+1}$ - ν^2- ν =

$\frac{(\nu^2+\nu)K!(N-K)!(N+2)!}{N!(N-K)!(K+2)!}$ - $(2\nu^2+\nu)\frac{N-K}{K+1}$ - ν^2- ν=

$= \frac{(\nu^2+\nu)(N+1)(N+2)}{(K+1)(K+2)}$ - $(2\nu^2+\nu)\frac{(N-K)(K+2)}{(K+1)(K+2)}$ - $\nu^2\frac{(K+1)(K+2)}{(K+1)(K+2)}$ - $\nu\frac{(K+1)(K+2)}{(K+1)(K+2)}$ =

$\frac{1}{(K+1)(K+2)}$ [$(\nu^2+\nu)(N^2+3N+2)$ - $(2\nu^2+\nu)(NK+2N-K^2-2K$ - $\nu^2(K^2+3K+2) - \nu(K^2+3K+2)$=

$\frac{1}{(K+1)(K+2)}$ [$\nu^2N^2+3\nu^2N+2\nu^2+\nu N^2+3\nu N+2\nu-2\nu^2NK-4\nu^2N+2\nu^2K^2+4\nu^2K-N\nu\kappa-2\nu N+\nu K^2+2K\nu-\nu^2K^2-3\nu^2K-2\nu^2-\nu K^2-3K\nu-2\nu$]=

$\frac{1}{(K+1)(K+2)}$ [ν^2N^2 - $\nu^2N + \nu N^2 + \nu N-2\nu^2NK + \nu^2K^2 + \nu^2K - \nu NK-K\nu$]=

$\frac{\nu}{(K+1)(K+2)}$ [$\nu N^2 - \nu N + N^2 + N -2\nu NK + \nu K^2 + \nu K - NK-K$]=

$\frac{\nu}{(K+1)(K+2)}$ [$\nu(N^2-2NK+K^2) - \nu N + N^2 + N - K + \nu K-NK$] =

$\frac{\nu}{(K+1)(K+2)}$ [$\nu(N-K)^2 + (N-K) + N(N-K) - \nu(N-K)$]= $\frac{\nu(N-K)[N-\nu+(N-K)\nu+1]}{(K+1)(K+2)}$

Άρα Var$Y$= $EY^2 - E^2\,Y = \frac{\nu(N-K)[N-\nu+(N-K)\nu+1]}{(K+1)(K+2)}$ - $\frac{\nu^2(N-K)^2}{(K+1)^2}$ =

$\frac{\nu(N-K)[N-\nu+(N-K)\nu+1](K+1)}{(K+1)^2(K+2)}$ - $\frac{\nu^2(N-K)^2(K+2)}{(K+1)^2(K+2)}$ =

$\frac{\nu(N-K)}{(K+1)^2(K+2)}$ [$N-\nu+N\nu-K\nu+1+NK-\nu K+N\nu K-K^2\nu+K-\nu(N-K)(K+2)$]=

$\frac{\nu(N-K)}{(K+1)^2(K+2)}$ [$N-\nu+N\nu-K\nu+1+NK-\nu K+N\nu K-K^2\nu+K-\nu NK-2\nu N+\nu K^2+2\nu K$]=

$\frac{\nu(N-K)}{(K+1)^2(K+2)}$ (N-ν-νN+1+NK+K)= $\frac{\nu(N-K)}{(K+1)^2(K+2)}$ N(K-ν+1) +(K-ν+1)=

$$\frac{\nu(N-K)(N+1)(K-\nu+1)}{(K+1)^2(K+2)}$$

Άρα και Var$X$= $\frac{\nu(N-K)(N+1)(K-\nu+1)}{(K+1)^2(K+2)}$

Ακόμη CV= $\frac{\sigma}{\mu} = \frac{\sqrt{\frac{\nu(N-K)(N+1)(K-\nu+1)}{(K+1)^2(K+2)}}}{\nu\frac{N+1}{K+1}} = \sqrt{\frac{(N-K)(K-\nu+1)}{\nu(N+1)(K+2)}}$

Συγκεντρωτικά οι τύποι των αριθμητικών μέτρων που υπολογίστηκαν για την **αρνητική υπεργεωμετρική κατανομή** είναι:

$$\boldsymbol{\mu = E}X\boldsymbol{= \nu\frac{N+1}{K+1}}$$

$$\boldsymbol{\sigma^2 = \mathrm{Var}}X\boldsymbol{= \frac{\nu(N-K)(N+1)(K-\nu+1)}{(K+1)^2(K+2)}}$$

$$\boldsymbol{CV = \frac{\sigma}{\mu} = \sqrt{\frac{(N-K)(K-\nu+1)}{\nu(N+1)(K+2)}}}$$

ΑΝΑΚΕΦΑΛΑΙΩΣΗ

Συνοψίζοντας, αναφέρθηκαν βασικές εισαγωγικές έννοιες της Στατιστικής και της Θεωρίας Πιθανοτήτων και υπολογίστηκαν τα αριθμητικά μέτρα της μέσης τιμής, της διασποράς και του συντελεστή μεταβλητότητας των κυριότερων διακριτών κατανομών. Αυτά τα αριθμητικά μέτρα είναι θεωρητικά αφού αναφέρονται στον πληθυσμό. Παρακάτω θα αναλύσουμε τους μεθόδους εκτίμησης αυτών των μέτρων μέσω δείγματος.

ΚΕΦΑΛΑΙΟ 2 «ΜΕΘΟΔΟΙ ΕΚΤΙΜΗΣΗΣ»

ΕΙΣΑΓΩΓΗ

Στο παρών κεφάλαιο θα περιγράψουμε τρόπους εκτίμησης των αριθμητικών περιγραφικών μέτρων που είδαμε. Θα γίνει εισαγωγή στον τομέα της Εκτιμητικής του κλάδου της Στατιστικής και θα γίνει ανάλυση της μεθόδου της Πιθανοφάνειας.

Θα αναφερθούμε στις εκτιμήτριες συναρτήσεις και στις έννοιες της μεροληψίας, της αποτελεσματικότητας και της συνέπειας για μικρό ή μεγάλο δείγμα. Επίσης θα εκτιμήσουμε τα περιγραφικά μέτρα της μέσης τιμής και της διασποράς και θα περιγράψουμε το πρόβλημα της εκτίμησης του συντελεστή μεταβλητότητας.

Πιο συγκεκριμένα θα τεθούν οι βάσεις για την εκτίμηση του συντελεστή μεταβλητότητας με τη χρήση της μεθόδου της Πιθανοφάνειας, και θα αναφερθούν οι χρήσιμες ιδιότητες των εκτιμητών με αυτή τη μέθοδο.

2.1 ΕΚΤΙΜΗΤΙΚΗ

Παρακάτω θα δώσουμε κάποιους ορισμούς απαραίτητους για την συνέχεια της μελέτης.

Ορισμός 2.1 **Στατιστικό** είναι ένα αριθμητικό περιγραφικό μέτρο που υπολογίζεται από το δείγμα.

Ορισμός 2.2 **Παράμετρος** είναι ένα αριθμητικό περιγραφικό μέτρο που υπολογίζεται από τον πληθυσμό.

Κύρια επιδίωξη της εκτιμητικής είναι να εκτιμηθούν οι παράμετροι της κατανομής κάποιου χαρακτηριστικού ενός πληθυσμού. Για παράδειγμα το ύψος των αντρών σε μία πόλη μπορεί να ακολουθεί κάποια γνωστή κατανομή $F(x;\theta)$ με παράμετρο $\theta=(\mu, \sigma^2)$. Για να εκτιμηθούν οι παράμετροι μ και σ^2 (μέση τιμή και διασπορά αντίστοιχα) θα πρέπει να γνωρίζουμε τις τιμές των χαρακτηριστικών αυτών σε κάποιες μονάδες του πληθυσμού. Είναι απαραίτητη δηλαδή η επιλογή δείγματος το οποίο θα πρέπει να είναι αντιπροσωπευτικό, δηλαδή να είναι παρμένο με τυχαίο τρόπο.

Ορισμός 2.3 **Στατιστική συνάρτηση** θα λέγεται κάθε συνάρτηση $T(X)=T(X_1,X_2,.....,X_n)$ των τυχαίων μεταβλητών του δείγματος $X_1,X_2,.....,X_n$ που δεν εξαρτάται από τις προς

εκτίμηση παραμέτρους. Προφανώς κάθε στατιστική συνάρτηση είναι και αυτή μία τυχαία μεταβλητή.

Για παράδειγμα γνωστές στατιστικές συναρτήσεις είναι :

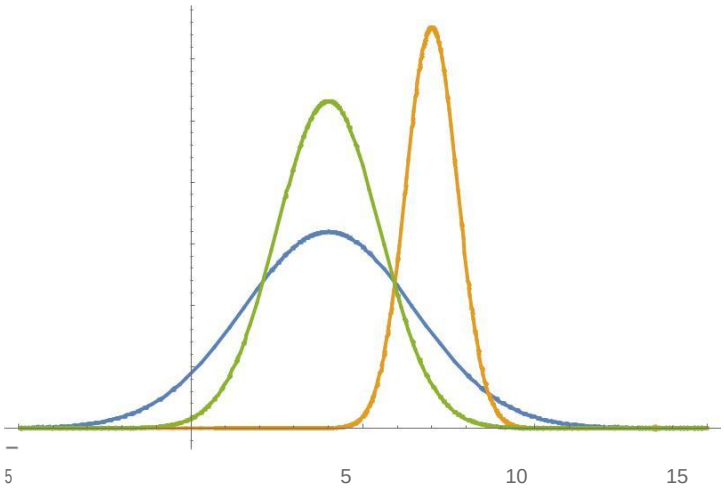
$m=\frac{1}{n}\sum_{i=1}^{n} xi$, ο δειγματικός μέσος και

$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(xi-m)^2$ η δειγματική διασπορά.

Ορισμός 2.4 **Εκτιμήτρια συνάρτηση** μιας παραμέτρου θ, θα καλείται μια στατιστική συνάρτηση $T(X_1,X_2,X_{3,\ldots\ldots,}X_n)$ η οποία χρησιμοποιείται για την εκτίμηση της θ.

Μία εκτιμήτρια συνάρτηση Τ μιας παραμετρικής συνάρτησης g(θ) ,για να θεωρηθεί ότι είναι «καλή» εκτιμήτρια της g(θ), θα πρέπει να έχει κάποια συγκεκριμένα χαρακτηριστικά όπως για παράδειγμα να παίρνει τιμές «πολύ κοντά» στην g(θ) με «μεγάλη» πιθανότητα. Αυτό μπορεί να γίνει απαιτώντας η τυχαία μεταβλητή Τ να έχει μέση τιμή ίση με g(θ) ή «σχεδόν» g(θ) και να έχει πολύ μικρή διασπορά(οι τιμές της τυχαίας μεταβλητής Τ να βρίσκονται «μαζεμένες» γύρω από τη μέση της τιμή).

Έτσι αν έχουμε τρεις διαφορετικές εκτιμήτριες T_1,T_2,T_3 (T_1= Πράσινη, T_2= Μπλε, T_3= Πορτοκαλί) για την μέση τιμή μ ενός πληθυσμού για την οποία γνωρίζουμε ότι μ=4 τότε από το παρακάτω γράφημα που παριστάνει τις συναρτήσεις πιθανότητας των εκτιμητριών έχουμε ότι



Γράφημα 1: Συναρτήσεις πιθανότητας των εκτιμητριών T_1, T_2, T_3

θα πρέπει να προτιμήσουμε την T_1 γιατί παίρνει τιμές πολύ κοντά στο 4 και έχει πιο μικρή διασπορά από την T_2 που παίρνει και αυτή τιμές γύρω από το 4, αλλά οι τιμές της μπορεί να διαφέρουν αρκετά από το μ γιατί έχει μεγάλη διασπορά. Η T_3 παίρνει τιμές που καμία σχέση δεν έχουν με την μ=4.

Ορισμός 2.5 Μία εκτιμήτρια συνάρτηση T της g(θ) θα καλείται **αμερόληπτη** εάν E(T) = E(T($X_1, X_2, \ldots\ldots\ldots, Xn$))= g(θ) ∀θ.

Ορισμός 2.6 Μία εκτιμήτρια συνάρτηση T της g(θ) θα καλείται **ασυμπτωτικά αμερόληπτη** εάν $\lim_{n\to\infty}$ E(T($X1, X2, \ldots\ldots\ldots\ldots, Xn$)) = g(θ).

Ορισμός 2.7 Έστω T μία εκτιμήτρια συνάρτηση της g(θ). Η ποσότητα

b(T)=E(T) – g(θ) καλείται **μεροληψία** της εκτιμήτριας T.

Πόρισμα 2.1 Η μεροληψία μιας αμερόληπτης εκτιμήτριας είναι 0.

Είναι φανερό ότι μας ενδιαφέρουν αμερόληπτες ή σχεδόν αμερόληπτες εκτιμήτριες γιατί αντίθετη περίπτωση θα έχουμε υπερεκτίμηση ή υποεκτίμηση της ζητούμενης παραμέτρου.

Πρόταση 2.1 Έστω $X_1, X_2, \ldots\ldots\ldots, X_n$ ένα τυχαίο δείγμα από οποιαδήποτε κατανομή

F(x;θ) με μέση τιμή μ=μ(θ) και διασπορά $\sigma^2 = \sigma^2(\theta)$.
Η στατιστική συνάρτηση $\mathbf{m= \frac{1}{n} \sum_{i=1}^{n} xi}$ (δειγματικός μέσος) είναι **αμερόληπτη εκτιμήτρια** της παραμέτρου μ= E(*X*i) (μέσης τιμής) της κατανομής F(x; θ) και έχει διασπορά
$\mathbf{Var(m) = \frac{\sigma^2}{n}}$.

Απόδειξη

$E(m)= E(\frac{1}{n} \sum_{i=1}^{n} xi) = \frac{1}{n}\ E\ (\sum_{i=1}^{n} xi) = \frac{1}{n} \sum_{i=1}^{n} E(xi)\ = \frac{1}{n} \sum_{i=1}^{n} \mu\ = \frac{n\mu}{n} = \mu.$

Επίσης $Var(m) =Var(\frac{1}{n} \sum_{i=1}^{n} xi) = \frac{1}{n^2} Var\ (\sum_{i=1}^{n} xi) = \frac{1}{n^2} \sum_{i=1}^{n} Var(xi) =$

$\frac{1}{n^2}\ n\ \sigma^2 =\ \frac{\sigma^2}{n}$.

Τέλος απόδειξης

Πρόταση 2.2 Έστω $X_1, X_2, \ldots\ldots, X_n$ ένα τυχαίο δείγμα από οποιαδήποτε κατανομή F(x; θ) με μέση τιμή μ= μ(θ) και διασπορά $\sigma^2 = \sigma^2(\theta)$.

Η στατιστική συνάρτηση $\mathbf{s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (xi - m)^2}$ είναι **αμερόληπτη εκτιμήτρια** της σ^2 , ενώ η $(s^*)^2 = \frac{1}{n} \sum_{i=1}^{n} (xi - m)^2$ είναι μεροληπτική.

Απόδειξη

Παρατηρούμε ότι $\sum_{\iota=1}^{n} (xi - m)^2 = \sum_{i=1}^{n} [(xi - \mu) - (m - \mu)]^2 =$

$\sum_{i=1}^{n} [(xi - \mu)^2 + (m - \mu)^2 - 2\,(xi - \mu)\,(m - \mu) =$

$\sum_{\iota=1}^{n} (xi - \mu)^2 + \sum_{\iota=1}^{n} (m - \mu)^2 - 2(m - \mu) \sum_{i=1}^{n} (xi - \mu) =$

$\sum_{\iota=1}^{n} (xi - \mu)^2 + \sum_{\iota=1}^{n} (m - \mu)^2 - 2(m - \mu)\, n\, (m-\mu) =$

$\sum_{\iota=1}^{n} (xi - \mu)^2 - n\, (m- \mu)^2$.

Επομένως έχουμε $E(s^2) = \frac{1}{n-1} E(\sum_{i=1}^{n} (xi - m)^2) = \frac{1}{n-1} E(\sum_{\iota=1}^{n} (xi - \mu)^2 - n\, (m - \mu)^2) =$

$\frac{1}{n-1} (\sum_{i=1}^{n} E(xi - \mu)^2 - nE\, (m - \mu)^2)$

Γνωρίζουμε όμως ότι $E(xi - \mu)^2 = \text{Var } xi = \sigma^2$ και $E\,(m - \mu)^2 = \text{Var } m = \frac{\sigma^2}{n}$,επομένως

$E(s^2) = \frac{1}{n-1} \left(n\, \sigma^2 - n\, \frac{\sigma^2}{n}\right) = \frac{1}{n-1} \sigma^2\, (n-1) = \sigma^2$.

Τέλος για την $(s^*)^2$ έχουμε $E((s^*)^2) = E(\frac{1}{n} \sum_{i=1}^{n} (xi - m)^2) =$

$\frac{1}{n} [\, E(\sum_{i=1}^{n} (xi^2) - n\, m^2] = \frac{1}{n} [\sum_{i=1}^{n} (Exi^2) - n\, E(m^2)] = \frac{1}{n} [n\sigma^2 + n\mu^2 - n\, E(m^2)] =$

$\sigma^2 + \mu^2 - E(m^2)$.

Όμως $\text{Var } m = \frac{\sigma^2}{n} \leftrightarrow E(m^2) - E^2(m) = \frac{\sigma^2}{n} \leftrightarrow E(m^2) = \frac{\sigma^2}{n} + \mu^2$.

Συνεπώς από τις δύο τελευταίες σχέσεις $E((s^*)^2) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \frac{n-1}{n} \neq \sigma^2$.

Τέλος απόδειξης

Όπως έχει αναφερθεί για να θεωρηθεί η στατική συνάρτηση T ότι είναι καλή «καλή» εκτιμήτρια της g(θ),θα πρέπει να έχει μέση τιμή g(θ) ή «σχεδόν» g(θ) και να έχει μικρή διασπορά. Όμως για μία παραμετρική συνάρτηση g(θ) δύναται να βρεθούν πολλές αμερόληπτες εκτιμήτριες. Τότε θα θεωρήσουμε «καλύτερη» αυτή που έχει την μικρότερη διασπορά. Συνεπώς προχωράμε στον παρακάτω ορισμό.

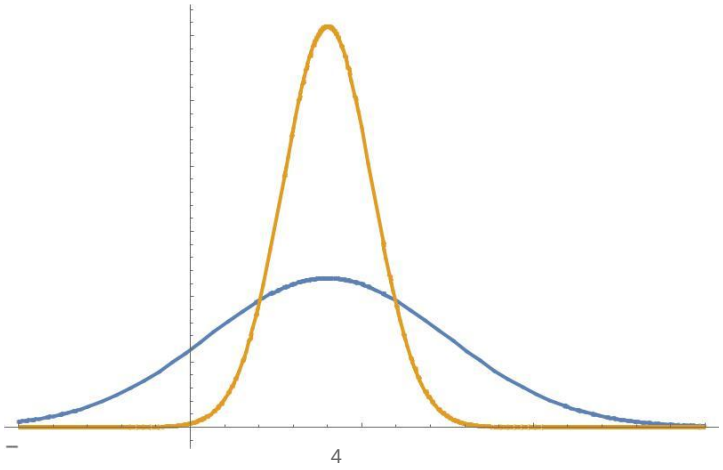
Ορισμός 2.8 Έστω T_1, T_2 δύο αμερόληπτες εκτιμήτριες της g(θ). Η T_1 θα καλείται **αποτελεσματικότερη** της T_2 αν ισχύει ότι

$Var(T_1) < Var(T_2)$.

Ορισμός 2.9 **Άριστη εκτιμήτρια ή αμερόληπτη εκτιμήτρια ελαχίστης διασποράς** της παραμετρικής συνάρτησης g(θ) λέγεται η αμερόληπτη εκτιμήτρια που έχει την μικρότερη διασπορά μεταξύ όλων των αμερόληπτων εκτιμητριών της g(θ).

Συνεπώς για τη βέλτιστη επιλογή εκτιμήτριας προκειμένου να εκτιμήσουμε μία παραμετρική συνάρτηση g(θ), πρέπει να επιλέξουμε μία άριστη εκτιμήτρια ή τουλάχιστον μία εκτιμήτρια που να είναι ασυμπτωτικά(για μεγάλο n) άριστη.

Αναφέραμε λοιπόν ότι μεταξύ αμερόληπτων εκτιμητριών επιλέγουμε αυτήν με τη μικρότερη διασπορά. Υπάρχουν όμως περιπτώσεις όπου έχουμε μη αμερόληπτες εκτιμήτριες ,με μικρή μεροληψία, που έχουν μικρότερη διασπορά από πολλές αμερόληπτες. Για παράδειγμα στο παρακάτω γράφημα όπου έχουμε τις συναρτήσεις πιθανότητας των εκτιμητριών T_1(μπλέ), T_2 (πορτοκαλί)



Γράφημα 2: Συναρτήσεις πιθανότητας των εκτιμητριών T_1, T_2

παρατηρούμε ότι η T_2 δεν είναι αμερόληπτη, καθώς έχει μέση τιμή λίγο παραπάνω από μ=4, ενώ η T_1 είναι αμερόληπτη καθώς έχει μέση τιμή ίση με 4. Ωστόσο η T_2 έχει πολύ μικρότερη διασπορά οπότε ίσως θεωρηθεί «καλύτερη» εκτιμήτρια γιατί παίρνει τιμές «πολύ κοντά» στο μ=4 με «μεγάλη» πιθανότητα.

Συνεπώς υπάρχει η ανάγκη για την δημιουργία μιας ποσότητας που θα μας εξασφαλίζει ποια εκτιμήτρια είναι καλύτερη. Οπότε συνεχίζουμε στον παρακάτω ορισμό.

Ορισμός 2.10 Έστω μια εκτιμήτρια $T=T(X_1,X_2,\ldots\ldots,X_n)$ μίας παραμέτρου g(θ). Η ποσότητα $mse(T)= E((T-g(\theta)^2)$ καλείται **μέσο τετραγωνικό σφάλμα** της T από την g(θ).

Επομένως βέλτιστη εκτιμήτρια μεταξύ εκτιμητριών μιας παραμέτρου θα θεωρούμε αυτή με το μικρότερο μέσο τετραγωνικό σφάλμα.

Πρόταση 2.3 Ισχύει $mse(T)=Var(T) + b^2(T)$.

Απόδειξη

$mse(T)= E((T-g(\theta)^2)= E[((T-E(T)) + (E(T)-g(\theta)))^2]= E[(T-E(T))^2 + (E(T)-g(\theta))^2 + 2(T-E(T))(E(T)-g(\theta))]= E((T-E(T))^2 + E((E(T)-g(\theta))^2) + 2E((T-E(T))(E(T)-g(\theta))$.

Όμως $E((T-E(T))(E(T)-g(\theta))=E(TE(T)-Tg(\theta)-E(T)^2+E(T)g(\theta))=E^2(T)-E(T)g(\theta)–E^2(T)+E(T)g(\theta)=0$

Συνεπώς $mse(T)= E((T-E(T))^2 + E((E(T)-g(\theta))^2)=Var(T) + (E(T)-g(\theta))^2 =Var(T) + b^2(T)$.

Τέλος απόδειξης

Πόρισμα 2.2 Ένας αμερόληπτος εκτιμητής έχει μέσο τετραγωνικό σφάλμα ίσο με τη διασπορά του.

Παρακάτω δίνουμε τον ορισμό για αποτελεσματικότερη εκτιμήτρια για τυχαίες εκτιμήτριες.

Ορισμός 2.11 Έστω T_1, T_2 δύο εκτιμήτριες της παραμετρικής συνάρτησης g(θ). Η T_1 καλείται **αποτελεσματικότερη** της T_2 αν ισχύει mse (T_1) < mse(T_2).

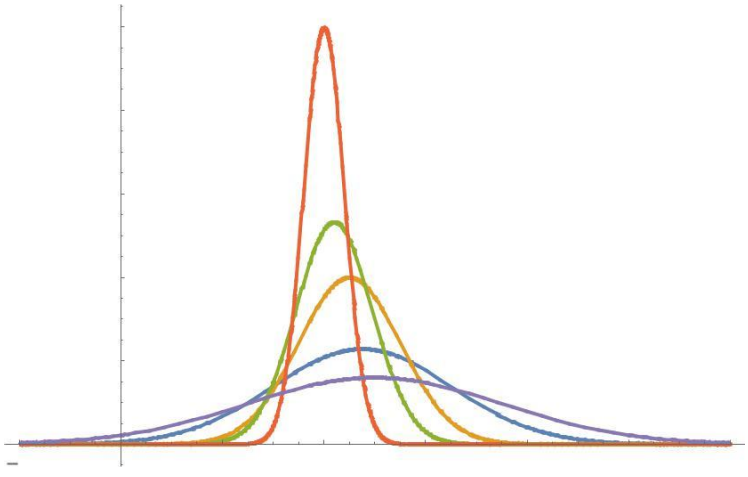
Αν οι εκτιμήτριες T_1,T_2 είναι αμερόληπτες τότε ο παραπάνω ορισμός είναι ισοδύναμος με τον ορισμό που αφορούσε αμερόληπτες εκτιμήτριες.

Ορισμός 2.12 **Σχετική αποτελεσματικότητα** μιας εκτιμήτριας T_1 σε σχέση με την T_2 ορίζεται το πηλίκο $\frac{\text{mse}(T2)}{\text{mse}(T1)}$.

Έστω $X_1,X_2,\ldots\ldots,X_n$ ένα τυχαίο δείγμα από μια κατανομή F(x; θ) και $T_n=T_n$ $(X_1,X_2,\ldots\ldots,X_n)$ μία στατιστική συνάρτηση που χρησιμοποιείται για την εκτίμηση της παραμετρικής συνάρτησης g(θ). Όπως έχουμε αναφέρει, θα θέλουμε η T_n να παίρνει τιμές «κοντά» στην g(θ) με «μεγάλη» πιθανότητα. Επίσης, αν πάρουμε πολύ μεγάλο δείγμα (θεωρητικά άπειρο) είναι φυσικό να ζητάμε καλύτερη ακρίβεια στην εκτίμηση της g(θ) από την T_n. Με άλλα λόγια μια εκτιμήτρια θα θεωρείται «καλή» (εκτός του ότι θέλουμε mse(T_n) να είναι «μικρό») και αν ,αυξάνοντας το μέγεθος του δείγματος, γίνεται ακριβέστερη ως προς την εκτίμηση της g(θ). Οπότε προχωράμε στον παρακάτω ορισμό.

Ορισμός 2.13 Μία εκτιμήτρια $T_n=T_n$ $(X_1,X_2,\ldots\ldots,X_n)$ μιας παραμέτρου g(θ) θα καλείται συνεπής αν ισχύει ότι $\lim_{n\to\infty} P(|Tn - g(\theta)| < \varepsilon) = 1$, $\forall \varepsilon > 0$.

Στο παρακάτω γράφημα έχουμε ένα παράδειγμα συνεπούς συνάρτησης όπου φαίνεται ότι καθώς μεγαλώνει το δείγμα η εκτιμήτρια γίνεται ακριβέστερη.



Γράφημα 3: Συνάρτηση πιθανότητας συνεπούς εκτιμήτριας

Σύμφωνα με τον παραπάνω ορισμό μία εκτιμήτρια $T_n=T_n(X_1,X_2,\ldots\ldots,X_n)$μιας παραμέτρου g(θ) θα είναι συνεπής αν λαμβάνοντας ένα πολύ μεγάλο δείγμα (ασυμπτωτικά) παίρνει τιμές σε μια οσοδήποτε μικρή περιοχή (g(θ)-ε,g(θ) + ε) γύρω από το g(θ) με πιθανότητα 1.Επειδή όμως ο έλεγχος δεν είναι εύκολο χρησιμοποιούμε την επόμενη πρόταση.

Πρόταση 2.4 Μία εκτιμήτρια $T_n=T_n(X_1,X_2,\ldots\ldots,X_n)$ μιας παραμέτρου g(θ) θα είναι **συνεπής** αν ισχύουν οι παρακάτω συνθήκες:

i) $\lim_{n\to\infty} E(Tn) = g(\theta)$ (δηλαδή $\lim_{n\to\infty} b(Tn) =0$)

ii) $\lim_{n\to\infty} Var(Tn) = 0$.

ή ισοδύναμα $\lim_{n\to\infty} mse(Tn) = \lim_{n\to\infty} Var(Tn) + b^2(Tn) =0$.

Απόδειξη

Από την ανισότητα Markov για την τυχαία μεταβλητή $(T_n - g(\theta))^2$ και α = ε έχουμε

$P((T_n - g(\theta)^2) \geq \varepsilon^2) \leq \frac{E[(Tn-g(\theta))^2]}{\varepsilon^2}$ $\forall \varepsilon > 0$ ή ισοδύναμα $P(|T_n - g(\theta)|\geq \varepsilon) \leq \frac{mse(Tn)}{\varepsilon^2}$.

Όμως $\lim_{n\to\infty} mse(Tn)=0$,άρα από κριτήριο παρεμβολής

Συνεπώς $\lim_{n\to\infty} P(|Tn - g(\theta)| < \varepsilon) =1$, ∀ε>0, άρα η T_n είναι συνεπής εκτιμήτρια της g(θ).

Τέλος απόδειξης

Πρόταση 2.5 Ισχύει η ανισότητα $P(|X|\geq \alpha) \leq \frac{E|X|}{\alpha}$ με α>0,γνωστή ως ανισότητα Markov.

Απόδειξη

Θεωρούμε την συνάρτηση $I_E = \begin{cases} 1, & αν\ το\ γεγονός\ E\ συμβαίνει \\ 0, & αν\ το\ γεγονός\ E\ δεν\ συμβαίνει \end{cases}$

και E το γεγονός $|x| \geq \alpha$.

Τότε $I_{(|x|\geq \alpha)} = \begin{cases} 1, αν\ |x| \geq \alpha \\ 0, & αν |x| < \alpha \end{cases}$

Τότε α $I_{(|x|\geq \alpha)} \leq |x|$, άρα $E(\alpha I_{(|x|\geq \alpha)}) \leq E(|X|)$.

Όμως $E(\alpha I_{(|x|\geq \alpha)}) = \alpha E(I_{(|x|\geq \alpha)}) = \alpha P(|X|\geq \alpha)$.

Από τις δύο τελευταίες σχέσεις προκύπτει ότι $\alpha P(|X|\geq \alpha) \leq E|X|$,δηλαδή

$P(|X|\geq \alpha) \leq \frac{E|X|}{\alpha}$,αφού α>0.

Τέλος απόδειξης

2.2 Ο ΔΕΙΓΜΑΤΙΚΟΣ ΕΚΤΙΜΗΤΗΣ ΤΟΥ CV

Όπως αναφέρθηκε στα προηγούμενα ο δειγματικός μέσος και η δειγματική διασπορά (με n-1 στον παρονομαστή) είναι αμερόληπτοι εκτιμητές της μέσης τιμής και της διασποράς του πληθυσμού αντίστοιχα. Ωστόσο η δειγματική τυπική απόκλιση δεν είναι αμερόληπτος εκτιμητής της τυπικής απόκλισης του πληθυσμού καθώς $E(s^2) = \sigma^2 \rightarrow \sqrt{E(s^2)} = \sigma$ και $E(s^2) \neq E^2(s)$. Συνεπώς $E(s) \neq \sigma$.

Δεδομένου ότι ο συντελεστής μεταβλητότητας εξαρτάται από την τυπική απόκλιση, θα ήταν πολύ δύσκολο να έχουμε αμερόληπτο συντελεστή μεταβλητότητας ενώ η τυπική απόκλιση είναι μεροληπτική. Οι συνθήκες γίνονται ακόμη πιο δύσκολες αν αναλογιστούμε και ότι $E(\frac{s}{m}) \neq \frac{Es}{Em}$. Συνεπώς έχουμε $E(CVdeig.)=E(\frac{s}{m}) \neq \frac{Es}{Em} \neq \frac{\sigma}{\mu} = CV$.

Αν και η τελευταία σχέση δεν αποδεικνύει ότι $E(CVdeig.) \neq CV$ θα δούμε παρακάτω κάποια παραδείγματα που δείχνουν το πρόβλημα της εκτίμησης του CV του πληθυσμού από τον δειγματικό CVdeig. Επίσης αναφέρουμε ,χωρίς να δώσουμε την απόδειξη, ότι από πρόσφατες αναφορές (Herve Abdi 2010) «σχεδόν» αμερόληπτος εκτιμητής του συντελεστή μεταβλητότητας και μόνο για κανονική κατανομή δίνεται από τον τύπο $CV^* = (1+\frac{1}{4n})$ CVdeig. , όπου CV^* ο «σχεδόν» αμερόληπτος εκτιμητής του CV και n το πλήθος του δείγματος ενώ η μεροληψία στην εκτίμηση του $\theta=CV^2$ από τον $\theta^*=CV^2$ deig. δίνεται από τον τύπο $Bias(\theta^*)=\frac{\theta^{\frac{3}{2}}}{n}(3\theta^{\frac{1}{2}} - \gamma_1)$ (Robert Breunig 2001). Η τελευταία σχέση ισχύει υπολογίζοντας

την μεροληψία του θ^* μέχρι $O(n^{-1})$ και λαμβάνοντας υπόψη μόνο τις πρώτες 6 κεντρικές ροπές τις οποίες θεωρούμε πεπερασμένες.

Παράδειγμα «κακής» εκτίμησης του CV από CVdeig. για κατανομή Bernoulli

Για την κατανομή Bernoulli είδαμε ότι $CV=\sqrt{\frac{1}{p}-1} \rightarrow CV^2=\frac{1-p}{P}$.

Συνεπώς ισχύει $\begin{cases} 0 \leq CV \leq 1, \ \alpha\nu \ p \geq \frac{1}{2} \\ 1 \leq CV \leq \infty, \ \alpha\nu \ p \leq \frac{1}{2} \end{cases}$

Θεωρούμε $X_1, X_2, \ldots\ldots, X_n$ ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή Bernoulli με παράμετρο $p \geq \frac{1}{2}$. Σύμφωνα με τα παραπάνω θα πρέπει $0 \leq CV \leq 1$ και συνεπώς αν ο CVdeig. είναι «καλός» εκτιμητής θα έπρεπε να ισχύει και $0 \leq CVdeig. \leq 1$.

Θα υπολογίσουμε την πιθανότητα να έχουμε «κακή» εκτίμηση του CV, από τον CVdeig., δηλαδή για $p=\frac{1}{2}$ την P(CVdeig >1).

Είναι $P(CVdeig >1)= P(CV^2 deig >1)= P(\frac{s^2}{m^2} >1)$.

Όμως $s^2 = \frac{\sigma^2 n}{n-1} = \frac{p(1-p)n}{n-1}$.

Από τις δύο τελευταίες σχέσεις έχουμε $P(CVdeig >1)= P(\frac{p(1-p)n}{(n-1)m^2} >1)$.**(Σχέση 2.1).**

Από εκτίμηση με τη μέθοδο πιθανοφάνειας (βλέπε κεφάλαιο 3) m= p.**(Σχέση 2.2).**

Συνεπώς από σχέσεις (2.1),(2.2) έχουμε $P(CVdeig >1)= P(\frac{n}{n-1}\frac{1-m}{m} >1)=$

$P(n-nm>nm-m)=P(n> m(2n-1))=P(n >\frac{\sum_{i=1}^{n} xi}{n}(2n-1)) =P(\sum_{i=1}^{n} xi < \frac{n^2}{2n-1}) =$

$\sum_{k=0}^{[\frac{n^2}{2n-1}]}\binom{n}{k} p^k(1-p)^{n-k}$ καθώς $\sum_{i=1}^{n} Xi \sim B(n, p)$ αφού $Xi \sim B(1,p)$.

([α] το ακέραιο μέρος του α)

Στον παρακάτω πίνακα βλέπουμε τις πιθανότητες «κακής» εκτίμηση του CV ,για διάφορα δείγματα και για διάφορες παραμέτρους p με p≥ 0.6.

p =	0.6	0.7	0.8	0.9
n=10	0.3669	0.1503	0.0328	0.0016
n=20	0.2447	0.048	0.0026	0.0000
n=30	0.1754	0.0169	0.0002	0.0000
n=40	0.1298	0.0063	0.0000	0.0000
n=50	0.0978	0.0024	0.0000	0.0000

Πίνακας 2: Πιθανότητες «κακής» εκτίμησης του CV στη Bernoulli

Παρατηρούμε ότι η μεγαλύτερη πιθανότητα συναντάται όταν n=10 και p=0.6, και είναι «αρκετά μεγάλη» με τιμή 0.3669 κάτι που δείχνει ότι ο CVdeig. δεν είναι «καλός» εκτιμητής.

Παράδειγμα «κακής» εκτίμησης του CV από CVdeig. για Ομοιόμορφη διακριτή κατανομή.

Θεωρούμε την τυχαία μεταβλητή *X* που ακολουθεί την Ομοιόμορφη διακριτή κατανομή στο [0, u-1]. Τότε από κεφάλαιο 1, η συνάρτηση πιθανότητας θα είναι $p(x)=\frac{1}{u}$, x=0,1,……,u-1.

Ακόμη CV= $\sqrt{\frac{u+1}{3(u-1)}}$.

Ισχύει η διπλή ανισότητα $\frac{\sqrt{3}}{3}$ < CV ≤ 1 , καθώς $\frac{\sqrt{3}}{3} < \sqrt{\frac{u+1}{3(u-1)}} \leq 1 \leftrightarrow$

$\frac{1}{3} < \frac{u+1}{3(u-1)} \leq 1 \leftrightarrow$ 3u-3<3u+3 και u+1≤ 3u-3 ↔ -3<3 και u≥2 που ισχύουν.

Συνεπώς αν ο CVdeig. είναι «καλός» εκτιμητής του CV του πληθυσμού θα πρέπει να ισχύει πάλι η διπλή ανισότητα $\frac{\sqrt{3}}{3} \cong 0.57$< CVdeig. ≤ 1.

Ο παρακάτω πίνακας δείχνει τις τιμές του CVdeig. για διάφορα δείγματα πλήθους 4 με επανάθεση από την διακριτή Ομοιόμορφη κατανομή.

$\sum_{i=1}^{4} xi$	Δείγμα	Αριθμός παρόμοιων δειγμάτων	CVdeig.
0	(0,0,0,0)	1	-
1	(1,0,0,0)	4	2
2	(2,0,0,0)	4	2
2	(1,1,0,0)	6	1.15
3	(2,1,0,0)	12	1.28
3	(1,1,1,0)	4	0.67
4	(2,2,0,0)	6	1.15
4	(2,1,1,0)	12	0.82
4	(1,1,1,1)	1	0
5	(2,2,1,0)	12	0.77
5	(2,1,1,1)	4	0.4
6	(2,2,2,0)	4	0.67
6	(2,2,1,1)	6	0.38
7	(2,2,2,1)	4	0.29
8	(2,2,2,2)	1	0

Πίνακας 3:Τιμές του δειγματικού CV στην Ομοιόμορφη διακριτή

Παρατηρούμε ότι εκτός του διαστήματος (0.57,1] βρίσκονται οι τιμές 2, 1.15, 1.28, 0, 0.4, 0.38, 0.29, που αντιστοιχούν σε (4+4) +(6+6)+12+(1+1)+4+6+4 = 48 περιπτώσεις δειγμάτων από τα 81 συνολικά.

Συνεπώς η πιθανότητα «κακής» εκτίμησης του CV θα είναι

P [(CVdeig.>1) ∪ (CVdeig.$\leq \frac{\sqrt{3}}{3}$)]= $\frac{48}{81}$ = 59% που είναι «αρκετά μεγάλη» κάτι που δείχνει ότι ο CVdeig. δεν είναι «καλός» εκτιμητής.

2.3 Η ΜΕΘΟΔΟΣ ΤΗΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Όπως αναφέρθηκε προηγουμένως η εκτίμηση του CV του πληθυσμού μέσω του CVdeig. ενδέχεται με «μεγάλη» πιθανότητα να μην είναι «καλή». Για αυτό θα χρησιμοποιήσουμε τη **μέθοδο της μέγιστης πιθανοφάνειας**. Πριν προχωρήσουμε στη περιγραφή της μεθόδου θα δώσουμε τον παρακάτω ορισμό.

Ορισμός 2.14 **Εκτιμητής σε σημείο** μιας παραμέτρου θ, είναι μία συνάρτηση από την οποία μπορούμε να υπολογίσουμε την αριθμητική τιμή της παραμέτρου, χρησιμοποιώντας τις μετρήσεις του δείγματος. Ο απλός αριθμός που είναι το αποτέλεσμα του υπολογισμού, λέγεται **εκτιμητής σε σημείο** της παραμέτρου.

Ορισμός 2.15 **Εκτιμητής σε διάστημα** είναι ένας τύπος από τον οποίο μπορούμε να υπολογίσουμε ένα διάστημα στο οποίο να ανήκει η παράμετρος του πληθυσμού ,χρησιμοποιώντας τις μετρήσεις του δείγματος.

Ορισμός 2.16 Το διάστημα στο οποίο με κάποιο μεγάλο βαθμό βεβαιότητας βρίσκεται η παράμετρος ενός πληθυσμού λέγεται **διάστημα εμπιστοσύνης** (δ.ε.) και η πιθανότητα με την οποία η εκτιμώμενη παράμετρος, περιέχεται στο διάστημα εμπιστοσύνης, λέγεται **συντελεστής εμπιστοσύνης.** Όταν ο συντελεστής εμπιστοσύνης είναι 1-α τότε το διάστημα εμπιστοσύνης λέγεται **100(1-α)% διάστημα εμπιστοσύνης.**

Ορισμός 2.17 Η τυχαία μεταβλητή X (συνεχής) που έχει παραμέτρους μ,σ^2 και συνάρτηση πυκνότητας πιθανότητας (αντίστοιχη της συνάρτησης πιθανότητας για διακριτές)

$f(x)= \frac{1}{\sigma\sqrt{2\pi}} \quad e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ λέμε ότι ακολουθεί την **κανονική κατανομή** και γράφουμε $\boldsymbol{X}\sim$ **N(μ, σ^2)** .

Πρόταση 2.6 **Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ.)**

Αν οι τυχαίες μεταβλητές $X_1,X_2,\ldots\ldots,X_n$ είναι ανεξάρτητες και ισόνομες (ακολουθούν την ίδια κατανομή) με $EX_i = \mu$, και $VarX_i = \sigma^2$, i=1,2,….,n τότε η δειγματική μέση τιμή m , ακολουθεί ασυμπτωτικά την κανονική κατανομή $N(\mu, \frac{\sigma^2}{n})$ που σημαίνει ότι :

$\frac{(m-\mu)\sqrt{n}}{\sigma} \sim N(0,1)$ ή $\sum_{\iota=1}^{n} X\iota \sim N(n\mu, n\sigma^2)$.

Πρόταση 2.7 **Ένα 100(1-α)% δ.ε. για το μ του πληθυσμού όταν το n είναι μεγάλο και το σ^2 άγνωστο** είναι το :

$(m- z_{a/2} \frac{s}{\sqrt{n}}, m- z_{a/2} \frac{s}{\sqrt{n}})$ όπου $z_{a/2}$ η τιμή εκείνη για την οποία

$P(-z_{a/2} < \frac{m-\mu}{(\sigma/\sqrt{n})} < z_{a/2}) = 1-a.$

Η εκτίμηση παραμέτρων σε σημείο γίνεται με αναλυτικές ή πιθανοθεωρητικές μεθόδους. Οι αναλυτικές μέθοδοι είναι η **μέθοδος των ροπών** , η **μέθοδος της μέγιστης πιθανοφάνειας** και η **μέθοδος των ελαχίστων τετραγώνων**. Στη παρούσα μελέτη θα περιγράψουμε και θα χρησιμοποιήσουμε την **μέθοδο της μέγιστης πιθανοφάνειας.**

Ορισμός 2.18 Ονομάζεται **πιθανοφάνεια** και **συμβολίζεται** με L(θ) , η κατανομή του δείγματος, όταν θεωρείται συνάρτηση της άγνωστης παραμέτρου θ. Αν P(x) είναι η συνάρτηση πιθανότητας από όπου προέρχεται το δείγμα τότε $L(\theta)= \prod_{i=1}^{n} P(xi)$.

Ο **εκτιμητής μέγιστης πιθανοφάνειας** (Ε.Μ.Π.) της άγνωστης παραμέτρου θ βρίσκεται μεγιστοποιώντας τη συνάρτηση L(θ) ως προς θ. Το μέγιστο της συνάρτησης L(θ) βρίσκεται παραγωγίζοντας τη συνάρτηση L(θ) ως προς θ . Ισοδύναμα παραγωγίζοντας τον λογάριθμο της L(θ), In(L(θ)) ως προς θ, χάριν ευκολίας στις πράξεις και επειδή η συνάρτηση Inx είναι αύξουσα, συνεπώς παρουσιάζει μέγιστο στο ίδιο σημείο. Επιπλέον πρέπει εκτός του ότι $\frac{dIn(L(\theta))}{d\theta}(\theta)=0$ και $\frac{d^2In(L(\theta))}{d^2\theta}(\theta) <0$. (Θεωρείται γνωστή η διαδικασία εύρεσης μεγίστου με χρήση παραγώγων)

Στα προηγούμενα αναφέραμε τις έννοιες της μεροληψίας , της αποτελεσματικότητας για αμερόληπτες εκτιμήτριες (συγκρίνοντας τις διασπορές) και της αποτελεσματικότητας για τυχαίες εκτιμήτριες (συγκρίνοντας το μέσο τετραγωνικό σφάλμα). Αυτές είναι ιδιότητες εκτιμητών για σχετικά «μικρό» δείγμα. Όσον αφορά τις ιδιότητες εκτιμητών «μεγάλων» δειγμάτων , δηλαδή τις ασυμπτωτικές για $n\rightarrow \infty$ αναφερθήκαμε στην έννοια της συνέπειας. Παρακάτω δίνονται ακόμη δύο ιδιότητες εκτιμητών «μεγάλων» δειγμάτων της ασυμπτωτικής αμεροληψίας και της ασυμπτωτικής αποτελεσματικότητας.

Ορισμός 2.19 Ο εκτιμητής Tn είναι **ασυμπτωτικά αμερόληπτος** εκτιμητής της παραμετρικής συνάρτησης g(θ) αν η μεροληψία του τείνει στο 0, καθώς το $n\rightarrow \infty$, δηλαδή $\lim_{n\rightarrow\infty} E(Tn) - g(\theta) =0$.

Ορισμός 2.20 Ο εκτιμητής T_n είναι **ασυμπτωτικά αποτελεσματικός** εκτιμητής της παραμετρικής συνάρτησης g(θ) αν έχει πεπερασμένο μέσο ,πεπερασμένη διακύμανση ,είναι

συνεπής και η διακύμανση της ασυμπτωτικής του κατανομής είναι μικρότερη από την διακύμανση της ασυμπτωτικής κατανομής οποιουδήποτε άλλου συνεπή εκτιμητή.

Πρόταση 2.8 Έστω Tn ο εκτιμητής μέγιστης πιθανοφάνειας του θ και g μία τυχαία συνάρτηση. Τότε ο **g(Tn) είναι ο εκτιμητής μέγιστης πιθανοφάνειας του g(θ).**

Η Πρόταση 2.8 και η απόδειξη της βρίσκεται στο κεφάλαιο 7 του συγγράμματος «Θέματα Παραμετρικής Στατιστικής Συμπερασματολογίας» (Σταύρος Κουρούκλης) και είναι ιδιαίτερα χρήσιμη καθώς θα αποτελέσει τη βάση του κεφαλαίου 3 (βλέπε παρακάτω). Αυτή η πρόταση είναι συμβατή με την αρχή της αντικατάστασης δηλαδή η εύρεση του Ε.Μ.Π. του θ, Tn,δίνει αμέσως τον Ε.Μ.Π της g(θ) που είναι ο $g(T_n)$, για τυχαία συνάρτηση g.Επίσης από «Θέματα Παραμετρικής Στατιστικής Συμπερασματολογίας» (Σταύρος Κουρούκλης) έχουμε τα εξής συμπεράσματα για τους Ε.Μ.Π.:

1) Οι Ε.Μ.Π γενικά έχουν πολύ καλή συμπεριφορά όταν υπάρχει μεγάλο πλήθος δεδομένων. Συγκεκριμένα σε τυπικές περιπτώσεις είναι συνεπείς, ασυμπτωτικά αμερόληπτοι και ασυμπτωτικά αποτελεσματικοί.

2) Οι Ε.Μ.Π ,εν γένει, δεν είναι αμερόληπτοι , όταν όμως είναι συμπίπτουν με τον αποτελεσματικό εκτιμητή.

ΑΝΑΚΕΦΑΛΑΙΩΣΗ

Συνοψίζοντας στο Κεφάλαιο 2 περιγράψαμε τους εκτιμητές των αριθμητικών περιγραφικών μέτρων της μέσης τιμής, της διασποράς και του συντελεστή μεταβλητότητας. Είδαμε ότι ενώ οι εκτιμητές της μέσης τιμής και της διασποράς είναι αμερόληπτοι , ο εκτιμητής του συντελεστή μεταβλητότητας ενδέχεται να μην είναι αρκετά «καλός», οπότε θέσαμε τις βάσεις για την εκτίμηση του με τη μέθοδο της πιθανοφάνειας. Στο Κεφάλαιο 3 γίνεται η διαδικασία αυτή , εκτιμάται δηλαδή ο συντελεστής μεταβλητότητας με τη μέθοδο της πιθανοφάνειας για τις κυριότερες διακριτές κατανομές.

ΚΕΦΑΛΑΙΟ 3 «ΕΚΤΙΜΗΣΗ ΤΟΥ CV ΜΕ ΠΙΘΑΝΟΦΑΝΕΙΑ»

ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό θα χρησιμοποιηθεί η μέθοδος της μέγιστης πιθανοφάνειας για την εκτίμηση των παραμέτρων των διακριτών κατανομών που περιγράφτηκαν στο Κεφάλαιο 1. Έτσι η εκτίμηση αυτή θα έχει τις ιδιότητες των εκτιμητών μέγιστης πιθανοφάνειας που περιγράφτηκαν στο Κεφάλαιο 2.

Στη συνέχεια θα χρησιμοποιηθεί η πρόταση 2.8 η οποία μας δίνει την δυνατότητα να εκτιμήσουμε όχι μόνο την παράμετρο θ, αλλά την παραμετρική συνάρτηση g(θ) για οποιαδήποτε τυχαία συνάρτηση g.

Πιο συγκεκριμένα η τυχαία αυτή συνάρτηση g θα είναι στη παρούσα μελέτη ο συντελεστής μεταβλητότητας των διακριτών κατανομών του Κεφαλαίου 1. Έτσι θα έχουμε μία εκτίμηση του CV, την MLE(CV) που θα έχει όλες τις ιδιότητες των Ε.Μ.Π. και θα είναι αρκετά «καλύτερη» από την εκτίμηση CVdeig. που είδαμε ότι σε αρκετές περιπτώσεις ενδέχεται να μην είναι «καλή» .

3.1 MLE(CV) ΓΙΑ ΟΜΟΙΟΜΟΡΦΗ ΔΙΑΚΡΙΤΗ ΚΑΤΑΝΟΜΗ

Έστω το τυχαίο δείγμα $X=(X_1,X_2,\ldots\ldots,X_n)$ που ακολουθεί την ομοιόμορφη διακριτή κατανομή U(α, β). Η συνάρτηση πιθανότητας της X είναι η : $P(x) = \frac{1}{\beta-\alpha+1}$ x=α, α+1,α+2,……,β. Υποθέτουμε ότι τα α, β είναι άγνωστα και θέλουμε να τα εκτιμήσουμε από δείγμα $x=(x_1,x_2,\ldots\ldots..,x_n)$ και θ=(α, β).

Τότε έχουμε $L(\theta/x)= \prod_{i=1}^{n} P(xi) = \frac{1}{(\beta-\alpha+1)^n} I_{[\alpha,\infty)}(x_{(1)}) I_{(-\infty,\beta]}(x_{(n)})$ όπου

$x_{(1)} = \min\{ x_1,x_2,\ldots\ldots..,x_n \}$ και $x_{(n)} = \max\{ x_1,x_2,\ldots\ldots..,x_n \}$.

Στη συγκεκριμένη περίπτωση παρατηρούμε ότι η L(θ/x) μεγιστοποιείται όταν

1)τα α, β είναι τέτοια έτσι ώστε η διαφορά β-α να γίνει ελάχιστη και

2) $x_{(1)} \geq \alpha$ και $x_{(n)} \leq \beta$.

Από τις δύο τελευταίες σχέσεις προκύπτει ότι

MLE(α)= min{ $x_1,x_2,\ldots\ldots,x_n$ } και MLE(β)= max{ $x_1,x_2,\ldots\ldots,x_n$ }.

Είδαμε ότι CV=$\sqrt{\frac{(\beta-\alpha+1)^2-1}{3(a+b)^2}}$ →MLE(CV)=$\sqrt{\frac{[MLE(\beta)-MLE(\alpha)+1]^2-1}{3[MLE(\alpha)+MLE(\beta)]^2}}$=

$\sqrt{\frac{[\max\{x1,x2,\ldots.xn\}-\min\{x1,x2,\ldots..,xn\}+1]^2-1}{3[\min\{xn\}+\max\{xn\}]^2}}$

Συνεπώς για την **ομοιόμορφη διακριτή κατανομή U(α, β)**

$$\mathbf{MLE(CV)} = \sqrt{\frac{[\mathbf{\max\{x1,x2,\ldots.xn\}-\min\{x1,x2,\ldots..,xn\}+1]^2-1}}{\mathbf{3[\min\{xn\}+\max\{xn\}]^2}}} \quad .$$

3.2 MLE(CV) ΓΙΑ BERNOULLI ΚΑΤΑΝΟΜΗ

Έστω τυχαία μεταβλητή *X* ακολουθεί κατανομή Bernoulli ($\boldsymbol{X}$**~B(1,p)**).Τότε είδαμε ότι P(X=x)= p^x $(1\text{-}p)^{1\text{-}x}$, x=0,1 και 0< p <1. Συνεπώς L(p)= $\prod_{i=1}^{n} P(xi) = p^{\sum_{i=1}^{n} xi} (1-p)^{n-\sum_{i=1}^{n} xi}$. Από ιδιότητες λογαρίθμων

In (L(p))= In $(p^{\sum_{i=1}^{n} xi} (1-p)^{n-\sum_{i=1}^{n} xi})$ = $\sum_{i=1}^{n} xi$ In(p) + (n - $\sum_{i=1}^{n} xi$) In(1- p).

Συνεπώς $\frac{dIn(L(p))}{dp} = \frac{\sum_{i=1}^{n} xi}{p} - \frac{n-\sum_{i=1}^{n} xi}{1-p} = 0 \leftrightarrow \frac{\sum_{i=1}^{n} xi}{p} = \frac{n-\sum_{i=1}^{n} xi}{1-p} \leftrightarrow$

$\sum_{i=1}^{n} xi - p\sum_{i=1}^{n} xi = pn - p\sum_{i=1}^{n} xi \leftrightarrow p=\frac{\sum_{i=1}^{n} xi}{p} \leftrightarrow p=m$

Επίσης $\frac{d^2In(L(p))}{d^2p} = -\frac{\sum_{i=1}^{n} xi}{p^2} - \frac{n-\sum_{i=1}^{n} xi}{(1-p)^2} \rightarrow$

$\frac{d^2In(L(p))}{d^2p}(p) = -\frac{mn}{m^2} - \frac{n-nm}{(1-m)^2} = -\frac{n}{m} - \frac{n}{1-m} < 0$, αφού 1- m>0.

Άρα MLE(p)= m

Επίσης είδαμε ότι CV= $\sqrt{\frac{1}{p}-1}$ → MLE(CV)= $\sqrt{\frac{1}{MLE(p)}-1}$ = $\sqrt{\frac{1}{m}-1}$

Συνεπώς για την **κατανομή Bernoulli B(1, p)**

MLE(CV) = $\sqrt{\frac{1}{m}-1}$

3.3 MLE(CV) ΓΙΑ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Έστω τυχαία μεταβλητή *X* ακολουθεί κατανομή την διωνυμική κατανομή

(***X*~B(n,p)**).Τότε είδαμε ότι P(*X*=x)= $\binom{N}{x}$ p^x $(1-p)^{N-x}$, x=0,1,……,N, 0< p <1 .

Άρα L(p)= $\prod_{i=1}^{n} P(xi) = \prod_{i=1}^{n}\binom{N}{xi} p^{xi} (1-p)^{N-xi} = p^{\sum_{i=1}^{n} xi} (1-p)^{nN-\sum_{i=1}^{n} xi} \prod_{i=1}^{n}\binom{N}{xi}$ →

In(L(p)) = $\sum_{i=1}^{n} xi$ Inp +(nN- $\sum_{i=1}^{n} xi$) In(1-p) + In ($\prod_{i=1}^{n}\binom{N}{xi}$) →

$\frac{dIn(L(p))}{dp} = \frac{\sum_{i=1}^{n} xi}{p} - \frac{nN-\sum_{i=1}^{n} xi}{1-p}$ =0 ↔ $\frac{\sum_{i=1}^{n} xi}{p} = \frac{nN-\sum_{i=1}^{n} xi}{1-p}$ ↔

$\sum_{i=1}^{n} xi - p\sum_{i=1}^{n} xi = p\,nN - p\sum_{i=1}^{n} xi$ ↔ p= $\frac{\sum_{i=1}^{n} xi}{nN}$ ↔ p= $\frac{m}{N}$

Επίσης $\frac{d^2In(L(p))}{d^2p} = -\frac{\sum_{i=1}^{n} xi}{p^2} - \frac{nN-\sum_{i=1}^{n} xi}{(1-p)^2}$ →

$\frac{d^2In(L(p))}{d^2p}(p) = -\frac{mN^2n}{m^2} - \frac{nN-mn}{(1-\frac{m}{N})^2} = -\frac{N^2n}{m} - \frac{n(N-m)}{(1-\frac{m}{N})^2} = -\frac{N^2n}{m} - \frac{nN^2}{N-m} < 0$ αφού N>m.

Συνεπώς MLE(p) = $\frac{m}{N}$

Είδαμε ότι CV= $\sqrt{\frac{1}{N}\left(\frac{1}{p}-1\right)}$ → MLE (CV) = $\sqrt{\frac{1}{N}\left(\frac{1}{MLE(p)}-1\right)}$ →

MLE (CV) = $\sqrt{\frac{1}{N}\left(\frac{N}{m}-1\right)} = \sqrt{\frac{1}{m}-\frac{1}{N}}$

Συνεπώς για την **διωνυμική κατανομή B(n, p)**

MLE (CV) = $\sqrt{\frac{1}{m} - \frac{1}{N}}$

3.4 MLE(CV) ΓΙΑ ΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Η τυχαία μεταβλητή X με συνάρτηση πιθανότητας $P(X=x)=P(x)=p(1-p)^x$, x=0,1,2,…. ,όπου $0< p <1$ ακολουθεί τη γεωμετρική κατανομή.

Άρα $L(p)= \prod_{i=1}^{n} P(xi) = \prod_{i=1}^{n} p\,(1-p)^{xi} = p^n\,(1-p)^{\sum_{i=1}^{n} xi} \rightarrow$

$In(L(p)) = In\,(p^n\,(1-p)^{\sum_{i=1}^{n} xi}) = n\,Inp + \sum_{i=1}^{n} xi\,In(1-p) \rightarrow$

$\frac{dIn(L(p))}{dp} = \frac{n}{p} - \frac{\sum_{i=1}^{n} xi}{1-p} = 0 \leftrightarrow \frac{n}{p} = \frac{\sum_{i=1}^{n} xi}{1-p} \leftrightarrow n - np = p\sum_{i=1}^{n} xi \leftrightarrow$

$n= p(\sum_{i=1}^{n} xi + n) \leftrightarrow p= \frac{n}{\sum_{i=1}^{n} xi + n} \leftrightarrow p= \frac{1}{m+1}$.Επίσης $\frac{d^2In(L(p))}{d^2p} = - \frac{n}{p^2} - \frac{\sum_{i=1}^{n} xi}{(1-p)^2} < 0\ \forall p.$

Άρα $MLE(p)= \frac{1}{m+1}$.

Είδαμε ότι $CV= \frac{1}{\sqrt{1-p}} \rightarrow MLE\,(CV) = \frac{1}{\sqrt{1-MLE(p)}} = \frac{1}{\sqrt{1-\frac{1}{m+1}}} = \frac{1}{\sqrt{\frac{m+1-1}{m+1}}} =$

$\sqrt{\frac{m+1}{m}} = \sqrt{1+\frac{1}{m}}$.

Συνεπώς για την **γεωμετρική κατανομή**

MLE(CV)= $\sqrt{\mathbf{1+\frac{1}{m}}}$**.**

3.5 MLE(CV) ΓΙΑ ΚΑΤΑΝΟΜΗ POISSON ΚΑΤΑΝΟΜΗ

Η τυχαία μεταβλητή *X* που ακολουθεί την **κατανομή Poisson** με παράμετρο λ, (*X*~P(λ)) έχει συνάρτηση πιθανότητας : $P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}$, x=0,1,2,….., λ>0 (παράμετρος) .

Άρα $L(\lambda) = \prod_{i=1}^{n} P(xi) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{xi}}{xi!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} xi}}{\prod_{i=1}^{n} xi!} \rightarrow In(L(\lambda)) = -n\lambda + \sum_{i=1}^{n} xi\, In\lambda - In\prod_{i=1}^{n} xi! \rightarrow \frac{dIn(L(\lambda))}{d\lambda} = -n + \frac{\sum_{i=1}^{n} xi}{\lambda} = 0 \leftrightarrow \lambda = m.$

Επίσης $\frac{d^2 In(L(\lambda))}{d^2\lambda} = -\frac{\sum_{i=1}^{n} xi}{\lambda^2} < 0 \ \forall\ \lambda.$

Συνεπώς MLE(λ) =m.

Είδαμε ότι $CV = \frac{1}{\sqrt{\lambda}} \rightarrow MLE(CV) = \frac{1}{\sqrt{MLE(\lambda)}} = \frac{1}{\sqrt{m}}$.

Συνεπώς για την **κατανομή Poisson**

MLE(CV)= $\frac{1}{\sqrt{m}}$.

3.6 MLE(CV) ΓΙΑ ΑΡΝΗΤΙΚΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Όπως είδαμε η τυχαία μεταβλητή *X* έχει την **αρνητική διωνυμική κατανομή** αν έχει συνάρτηση πιθανότητας : $P(X=x) = p(x) = \binom{\nu+x-1}{x} p^{\nu} (1-p)^x$, x=0,1,….. ,όπου $\nu > 0$ και p παράμετρος με 0 < p <1.

Άρα $L(p) = \prod_{i=1}^{n} P(xi) = \prod_{i=1}^{n}[\binom{\nu+xi-1}{xi} p^{\nu} (1-p)^{xi}] = p^{n\nu} (1-p)^{\sum_{i=1}^{n} xi} \prod_{i=1}^{n}\binom{\nu+xi-1}{xi} \rightarrow$

$In(L(p)) = n\nu\, Inp + \sum_{i=1}^{n} xi\, In(1-p) + In\prod_{i=1}^{n}\binom{\nu+xi-1}{xi} \rightarrow$

$\frac{dIn(L(p))}{dp} = \frac{n\nu}{p} - \frac{\sum_{i=1}^{n} xi}{1-p} = 0 \leftrightarrow \frac{n\nu}{p} = \frac{\sum_{i=1}^{n} xi}{1-p} \leftrightarrow n\nu - n\nu p = p\sum_{i=1}^{n} xi \leftrightarrow$

$n\nu = p(\sum_{i=1}^{n} xi + n\nu) \leftrightarrow p = \frac{n\nu}{\sum_{i=1}^{n} xi + n\nu} \leftrightarrow p = \frac{\nu}{m+\nu}$.

Επίσης $\frac{d^2 In(L(p))}{d^2 p} = -\frac{nv}{p^2} - \frac{\sum_{i=1}^{n} xi}{(1-p)^2} < 0 \ \forall$ p.

Συνεπώς MLE(p) = $\frac{\nu}{m+\nu}$.

Είδαμε ότι CV= $\frac{1}{\sqrt{\nu(1-p)}}$ →MLE (CV) = $\frac{1}{\sqrt{\nu(1-MLE(p))}}$ → MLE (CV) = $\frac{1}{\sqrt{\nu(1-\frac{\nu}{m+\nu})}}$ =

$\frac{1}{\sqrt{\nu\frac{m+\nu-\nu}{m+\nu}}} = \sqrt{\frac{m+\nu}{\nu m}} = \sqrt{\frac{1}{\nu} + \frac{1}{m}}$

Συνεπώς για την **αρνητική διωνυμική κατανομή**

MLE(CV) = $\sqrt{\frac{1}{\nu} + \frac{1}{m}}$.

3.7 MLE(CV) ΓΙΑ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Η τυχαία μεταβλητή *X* που παίρνει τις τιμές 0,1,……,ν, και ακολουθεί την υπεργεωμετρική κατανομή είδαμε ότι έχει συνάρτηση πιθανότητας: P(*X*=x)= p(x) =$\frac{\binom{K}{x}\binom{N-K}{\nu-x}}{\binom{N}{\nu}}$

,0≤ x ≤ ν, όπου N, K, ν είναι ακέραιοι με 0≤ K ≤ N ,1≤ ν ≤ N.

Όσον αφορά την εκτίμηση με τη μέθοδο της πιθανοφάνειας ,θεωρούμε άγνωστο μόνο το K, το οποίο θα εκτιμήσουμε με τη βοήθεια των N και ν που θεωρούνται γνωστά και σταθερά.

Από Hanwen Zhang(2009) «A Note About Maximum Likelihood Estimator in Hyper geometric Distribution» έχουμε ότι :

MLE(K)= $\begin{cases} m\ \frac{N+1}{\nu} - 1 \text{ ή } m\ \frac{N+1}{\nu}, \alpha\nu\ m\ \frac{N+1}{\nu} \in \mathbb{Z} \\ \left[m\ \frac{N+1}{\nu}\right], \ \alpha\nu\ m\ \frac{N+1}{\nu} \notin \mathbb{Z} \end{cases}$

Από τα παραπάνω συμπεραίνουμε ότι μπορούμε να επιλέξουμε MLE(K)= $\left[m\ \frac{N+1}{\nu}\right]$

Είχαμε δει ότι CV=$\sqrt{\frac{(N-K)(N-\nu)}{\nu K(N-1)}}$ → MLE (CV) = $\sqrt{\frac{(N-MLE(K))(N-\nu)}{\nu MLE(K)(N-1)}}$ =

$$\sqrt{\frac{\left(\mathrm{N}-[\frac{m(N+1)}{\nu}]\right)(\mathrm{N}-\nu)}{\nu[\frac{m(N+1)}{\nu}](\mathrm{N}-1)}}\,.$$

Συνεπώς για την **υπεργεωμετρική κατανομή**

$$\mathbf{MLE(CV)=}\sqrt{\frac{\left(\mathbf{N}-[\frac{\boldsymbol{m(N+1)}}{\boldsymbol{\nu}}]\right)(\mathbf{N}-\boldsymbol{\nu})}{\boldsymbol{\nu}[\frac{\boldsymbol{m(N+1)}}{\boldsymbol{\nu}}](\mathbf{N}-\mathbf{1})}}\,\mathbf{.}$$

3.8 MLE(CV) ΓΙΑ ΑΡΝΗΤΙΚΗ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Έστω μια κάλπη που περιέχει K άσπρα και N-K μαύρα σφαιρίδια. Εξάγουμε διαδοχικά ,ένα προς ένα χωρίς επανάθεση τα σφαιρίδια μέχρι να πάρουμε το ν-οστό άσπρο σφαιρίδιο, με ν≤ K. Αν η τυχαία μεταβλητή που μετρά το πλήθος των απαιτούμενων εξαγωγών, τότε η *X* ακολουθεί την αρνητική υπεργεωμετρική κατανομή. Η συνάρτηση πιθανότητας της *X* είναι: $\mathrm{P}(X=\mathrm{x}) = \mathrm{p(x)}=\frac{\binom{\mathrm{K}}{\nu-1}\binom{\mathrm{N}-\mathrm{K}}{\mathrm{x}-\nu}}{\binom{\mathrm{N}}{\mathrm{x}-1}}\frac{\mathrm{K}-\nu+1}{\mathrm{N}-\mathrm{x}+1}$, x= ν, ν +1,…….,ν+N-K.

Για την εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας θεωρούμε γνωστά τα N, ν και εκτιμάμε το K.

Από Lei Zhang και William D.Johnson (2011) «Approximate Confidence Intervals for a Parameter of the Negative Hyper geometric Distribution» έχουμε ότι :

$\mathrm{MLE(K)}=\frac{(\nu-1)\mathrm{N}}{\mathrm{m}-1}$ (Theorem 3.2)

Είδαμε ότι $\mathrm{CV}=\sqrt{\frac{(\mathrm{N}-\mathrm{K})(\mathrm{K}-\nu+1)}{\nu(\mathrm{N}+1)(\mathrm{K}+2)}}$ → $\mathrm{MLE(CV)}=\sqrt{\frac{(\mathrm{N}-\mathrm{MLE(K)})(\mathrm{MLE(K)}-\nu+1)}{\nu(\mathrm{N}+1)(\mathrm{MLE(K)}+2)}}=$

$$\sqrt{\frac{\left(\mathrm{N}-\frac{(\nu-1)\mathrm{N}}{\mathrm{m}-1}\right)\left(\frac{(\nu-1)\mathrm{N}}{\mathrm{m}-1}-\nu+1\right)}{\nu(\mathrm{N}+1)\left(\frac{(\nu-1)\mathrm{N}}{\mathrm{m}-1}+2\right)}}$$

Συνεπώς για την **αρνητική υπεργεωμετρική κατανομή**

$$\textbf{MLE(CV)}=\sqrt{\frac{\left(\mathbf{N}-\frac{(\boldsymbol{\nu}-\mathbf{1})\mathbf{N}}{\mathbf{m}-\mathbf{1}}\right)\left(\frac{(\boldsymbol{\nu}-\mathbf{1})\mathbf{N}}{\mathbf{m}-\mathbf{1}}-\boldsymbol{\nu}+\mathbf{1}\right)}{\boldsymbol{\nu}(\mathbf{N}+\mathbf{1})\left(\frac{(\boldsymbol{\nu}-\mathbf{1})\mathbf{N}}{\mathbf{m}-\mathbf{1}}+\mathbf{2}\right)}}$$

ΑΝΑΚΕΦΑΛΑΙΩΣΗ

Συνοψίζοντας, στο παρών κεφάλαιο βρήκαμε τους εκτιμητές μέγιστης πιθανοφάνειας για τις παραμέτρους των κυριότερων διακριτών κατανομών και μέσω αυτών και της πρότασης 2.8 εκτιμήσαμε τον συντελεστή μεταβλητότητας για κάθε διακριτή κατανομή. Οι εκτιμητές αυτοί περιμένουμε να έχουν τις ιδιότητες της συνέπειας της ασυμπτωτικής αμεροληψίας και της αποτελεσματικότητας αφού είναι εκτιμητές μέγιστης πιθανοφάνειας. Συνεπώς αναμένουμε σύγκλιση της εκτίμησης στην πραγματική τιμή του συντελεστή μεταβλητότητας (του πληθυσμού) και γενικώς «καλά» αποτελέσματα (αμεροληψία, αποτελεσματικότητα) για «μεγάλο» δείγμα ($n\rightarrow\infty$). Τα θέματα αυτά μελετούνται στο επόμενο κεφάλαιο στο προγραμματιστικό περιβάλλον της R.

ΚΕΦΑΛΑΙΟ 4 «ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΣΤΗΝ R»

ΕΙΣΑΓΩΓΗ

Στο Κεφάλαιο 4 θα προσομοιώσουμε την θεωρία που αναπτύχθηκε στα προηγούμενα κεφάλαια. Θα επικεντρωθούμε στη δειγματοληψία με επανάθεση για τις διακριτές κατανομές που μελετήθηκαν στο Κεφάλαιο 1 και στην επιλογή του κατάλληλου μεγέθους δείγματος έτσι ώστε να έχουμε επιθυμητό απόλυτο σφάλμα προσέγγισης.

Ο χρήστης θα επιλέγει ένα απόλυτο σφάλμα προσέγγισης του CV(του πληθυσμού) από το CVdeig.(δειγματικό) και από το MLE(CV) (δειγματικό με τη μέθοδο της πιθανοφάνειας). Έπειτα με υπολογιστικές μεθόδους ο αλγόριθμος θα δίνει απάντηση για το τι δείγμα πρέπει να πάρουμε σε μορφή διαστήματος εμπιστοσύνης και για τις δύο περιπτώσεις.

Πιο συγκεκριμένα ο αλγόριθμος σε κάθε βήμα παράγει δείγματα με επανάθεση, με αυξανόμενο πλήθος (ξεκινώντας από 1) και υπολογίζει CV(που είναι σταθερό) , CV deig. και MLE(CV) .Έπειτα υπολογίζει τα απόλυτα σφάλματα και σταματάει όταν το απόλυτο σφάλμα γίνει για $1^{η}$ φορά. μικρότερο από αυτό που επέλεξε ο χρήστης.

4.1 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΟΜΟΙΟΜΟΡΦΗ ΚΑΤΑΝΟΜΗ

Σε αυτήν την ενότητα θα εξηγήσουμε με μεγάλη λεπτομέρεια το πρόγραμμα που δημιουργήθηκε για τις ανάγκες της συγκεκριμένης μελέτης προκειμένου να γίνει όσο το δυνατό κατανοητό από τον αναγνώστη.

Αρχικά απαραίτητες για το πρόγραμμα είναι οι παρακάτω 2 προκαταρκτικές συναρτήσεις:

1) f(x ,n) και 2)combin(n, k) .

Επεξήγηση της f(x ,n)

Η συνάρτηση f(x, n) δέχεται ένα διάνυσμα τιμών x , και ένα πραγματικό αριθμό n.

Ελέγχει από τα αριστερά προς τα δεξιά αν στις τιμές του διανύσματος υπάρχει μία που να είναι μικρότερη από τον αριθμό n. Αν υπάρχει επιστρέφει τη δυάδα (1 , x_o) όπου x_o η θέση του πρώτου αριθμού που βρέθηκε να είναι μικρότερος από το n ,και το 1 είναι ένδειξη ότι βρέθηκε. Αν δεν υπάρχει τέτοιος αριθμός τότε επιστρέφει τη δυάδα (0,Inf) .

Για παράδειγμα

```
> f(c(6,7,8,9,6,6,7,8,7,8,7,4),5)
[1]  1 12
```

Ενώ

```
> f(c(6,7,8,9,6,5),4)
[1]   0 Inf
```

Επεξήγηση της combin(n,k)

Η συνάρτηση αυτή υπολογίζει τους συνδυασμούς $\binom{n}{k}$. Για παράδειγμα

```
> combin(10,2)
[1] 45
```

Συνεχίζουμε με το κυρίως κομμάτι του κώδικα για ομοιόμορφη διακριτή κατανομή. Αρχικά πρέπει να βάλουμε τις παραμέτρους της κατανομής και το επιθυμητό σφάλμα προσέγγισης.

Επιλέγουμε α=1, β=100 και ERROR=0.0001.

Συνεπώς μέσω της εντολής K =c(a:b) το σύνολο από το οποίο θα γίνει η δειγματοληψία είναι το

```
> K
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 [91]  91  92  93  94  95  96  97  98  99 100
```

Πίνακας 4 : Σύνολο αναφοράς Ω στην Ομοιόμορφη διακριτή U(1,100)

Συνεχίζοντας ο αλγόριθμος σε κάθε βήμα παράγει δείγματα με δειγματοληψία με επανάθεση ξεκινώντας από μέγεθος 1 και αυξάνοντας συνεχώς το μέγεθος του δείγματος. Η παραπάνω παραγωγή δεδομένων γίνεται μέσω της εντολής

x=sample(K,i, prob=Prob, rep=T) όπου Prob η συνάρτηση πιθανότητας.

Πιο συγκεκριμένα κάθε στοιχείο x επιλέγεται με βάρος ίσο με P(X=x).

Έτσι εξασφαλίζουμε ότι τα δεδομένα ακολουθούν την εκάστοτε κατανομή.

Για κάθε δείγμα υπολογίζει τον πίνακα M, όπου στη $1^{η}$ στήλη έχει το CVdeig.,στην $2^{η}$ στήλη το CV(πληθυσμού) και στην $3^{η}$ το MLE(CV).

Στο παράδειγμα μας

```
> M
             [,1]      [,2]        [,3]
 [1,] 0.00000000 0.0000000 0.000000000
 [2,] 0.01321695 0.5716053 0.009345794
 [3,] 0.62230281 0.5716053 0.328886875
 [4,] 0.68717949 0.5716053 0.478893352
 [5,] 0.52033204 0.5716053 0.428098243
 [6,] 0.61736128 0.5716053 0.448447814
 [7,] 0.96652160 0.5716053 0.527834464
 [8,] 0.49828999 0.5716053 0.549019608
 [9,] 0.57720194 0.5716053 0.441149130
[10,] 0.70603731 0.5716053 0.560340791
[11,] 0.54013852 0.5716053 0.486441687
[12,] 0.63797279 0.5716053 0.447286183
[13,] 0.47377841 0.5716053 0.558068391
[14,] 0.67798712 0.5716053 0.516836707
[15,] 0.68360623 0.5716053 0.536078593
[16,] 0.57955189 0.5716053 0.475749355
[17,] 0.64682376 0.5716053 0.548452368
[18,] 0.41647647 0.5716053 0.570826328
[19,] 0.64133083 0.5716053 0.482026858
[20,] 0.63267399 0.5716053 0.571040241
[21,] 0.63338133 0.5716053 0.560172086
[22,] 0.40071090 0.5716053 0.388459905
[23,] 0.63963844 0.5716053 0.548738799
[24,] 0.50449968 0.5716053 0.476681804
[25,] 0.69435374 0.5716053 0.571488693
[26,] 0.58024914 0.5716053 0.538083677
[27,] 0.48459516 0.5716053 0.468092943
[28,] 0.59897065 0.5716053 0.571605347
[29,] 0.50575482 0.5716053 0.516836707
[30,] 0.70251378 0.5716053 0.560000000
[31,] 0.67395896 0.5716053 0.560340791
[32,] 0.68336278 0.5716053 0.559824432
[33,] 0.47873263 0.5716053 0.560000000
[34,] 0.60072652 0.5716053 0.560340791
[35,] 0.60486947 0.5716053 0.559824432
[36,] 0.57233544 0.5716053 0.549019608
[37,] 0.56999637 0.5716053 0.526872371
```

```
[38,] 0.69479793 0.5716053 0.571547607
[39,] 0.64867630 0.5716053 0.571547607
[40,] 0.58327144 0.5716053 0.571547607
[41,] 0.45934674 0.5716053 0.527834464
[42,] 0.58200395 0.5716053 0.571605347
[43,] 0.48217216 0.5716053 0.571605347
[44,] 0.48845108 0.5716053 0.527834464
[45,] 0.51628997 0.5716053 0.517407894
[46,] 0.51646700 0.5716053 0.571428571
[47,] 0.59078640 0.5716053 0.560172086
[48,] 0.57555058 0.5716053 0.560340791
[49,] 0.53197269 0.5716053 0.486441687
[50,] 0.53144131 0.5716053 0.560172086
[51,] 0.54793167 0.5716053 0.571605347
[52,] 0.53451559 0.5716053 0.560340791
[53,] 0.63320000 0.5716053 0.560340791
[54,] 0.64574648 0.5716053 0.571488693
[55,] 0.59500502 0.5716053 0.571547607
[56,] 0.59841326 0.5716053 0.571605347
[57,] 0.54813857 0.5716053 0.571605347
[58,] 0.67185184 0.5716053 0.560340791
[59,] 0.48567300 0.5716053 0.571547607
[60,] 0.59528333 0.5716053 0.507176208
[61,] 0.58047485 0.5716053 0.571547607
[62,] 0.48811383 0.5716053 0.506513877
[63,] 0.49136280 0.5716053 0.549019608
[64,] 0.63529033 0.5716053 0.571605347
[65,] 0.59716327 0.5716053 0.538461538
[66,] 0.55934156 0.5716053 0.571547607
[67,] 0.56193441 0.5716053 0.571605347
[68,] 0.50148053 0.5716053 0.571367203
[69,] 0.58302408 0.5716053 0.571547607
[70,] 0.52940983 0.5716053 0.495619832
[71,] 0.55302764 0.5716053 0.571605347
[72,] 0.60221784 0.5716053 0.548738799
[73,] 0.65073178 0.5716053 0.571605347
[74,] 0.54417594 0.5716053 0.571605347
[75,] 0.59634808 0.5716053 0.560340791
[76,] 0.55712832 0.5716053 0.571367203
```

```
 [77,] 0.66817117 0.5716053 0.571428571
 [78,] 0.51310034 0.5716053 0.571547607
 [79,] 0.60179542 0.5716053 0.560340791
 [80,] 0.58273085 0.5716053 0.571547607
 [81,] 0.61777289 0.5716053 0.571547607
 [82,] 0.58928034 0.5716053 0.560000000
 [83,] 0.53654908 0.5716053 0.571605347
 [84,] 0.56324634 0.5716053 0.571547607
 [85,] 0.48308696 0.5716053 0.538461538
 [86,] 0.63461693 0.5716053 0.571605347
 [87,] 0.67242388 0.5716053 0.571547607
 [88,] 0.50374360 0.5716053 0.560340791
 [89,] 0.63594124 0.5716053 0.571605347
 [90,] 0.62534988 0.5716053 0.560000000
 [91,] 0.53472891 0.5716053 0.571605347
 [92,] 0.61390130 0.5716053 0.571605347
 [93,] 0.50926593 0.5716053 0.571547607
 [94,] 0.67323849 0.5716053 0.571547607
 [95,] 0.53590509 0.5716053 0.559824432
 [96,] 0.59100684 0.5716053 0.571547607
 [97,] 0.55662105 0.5716053 0.571605347
 [98,] 0.66818841 0.5716053 0.571488693
 [99,] 0.51369503 0.5716053 0.571605347
[100,] 0.59442626 0.5716053 0.571547607
[101,] 0.64278648 0.5716053 0.571605347
[102,] 0.58251932 0.5716053 0.571428571
[103,] 0.58291152 0.5716053 0.571547607
[104,] 0.58047043 0.5716053 0.571605347
[105,] 0.59753724 0.5716053 0.571547607
[106,] 0.56624533 0.5716053 0.571605347
[107,] 0.60776409 0.5716053 0.571428571
[108,] 0.54638474 0.5716053 0.571605347
[109,] 0.57156795 0.5716053 0.560340791
```

Πίνακας 5: Συντελεστές μεταβλητότητας CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Ομοιόμορφη διακριτή U(1,100)

Ο αλγόριθμος σταματάει μόλις |CV- CVdeig.|<ERROR και |CV-MLE(CV)|<ERROR.

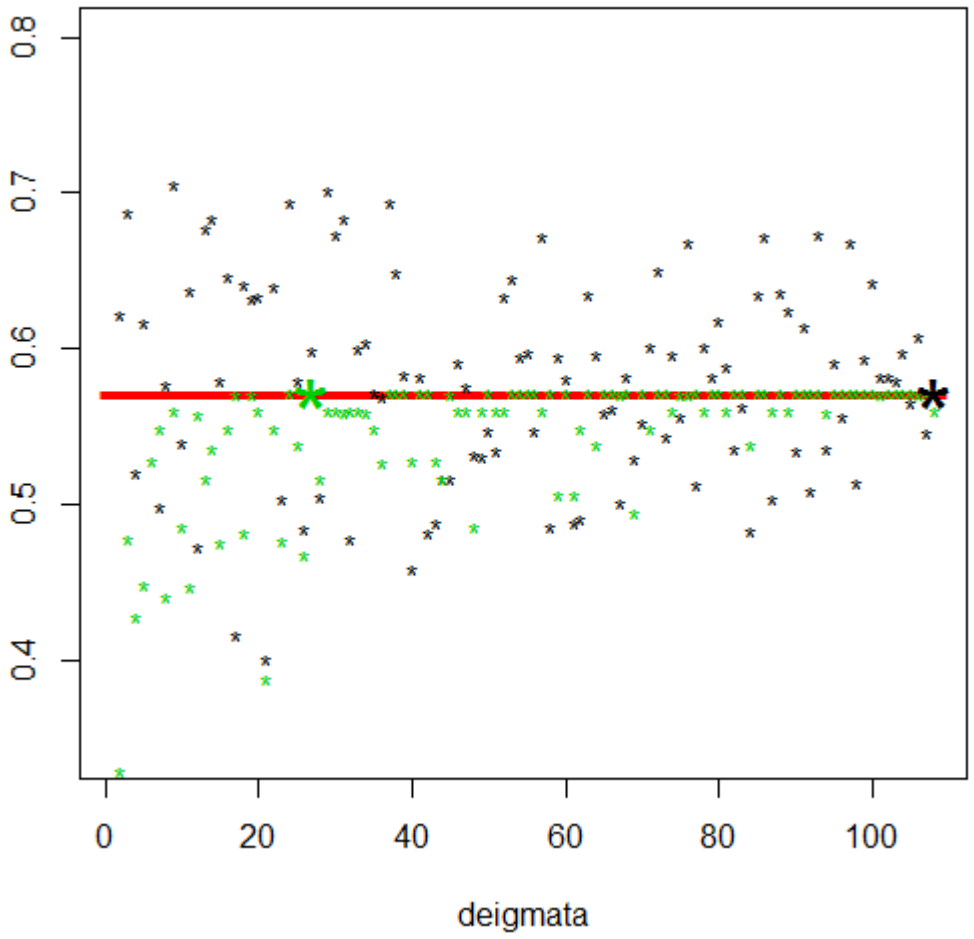
Η απάντηση για το πότε συμβαίνει αυτό γίνεται με την εντολή NCV και NMLECV αντίστοιχα. Στο παράδειγμα μας

```
> NCV
[1] 109
> NMLECV
[1] 28
```

Όντως αν παρατηρήσουμε την 109^{η} και την 28^{η} γραμμή θα δούμε ότι τα απόλυτα σφάλματα είναι μικρότερα του 0.0001 αντίστοιχα για κάθε περίπτωση.

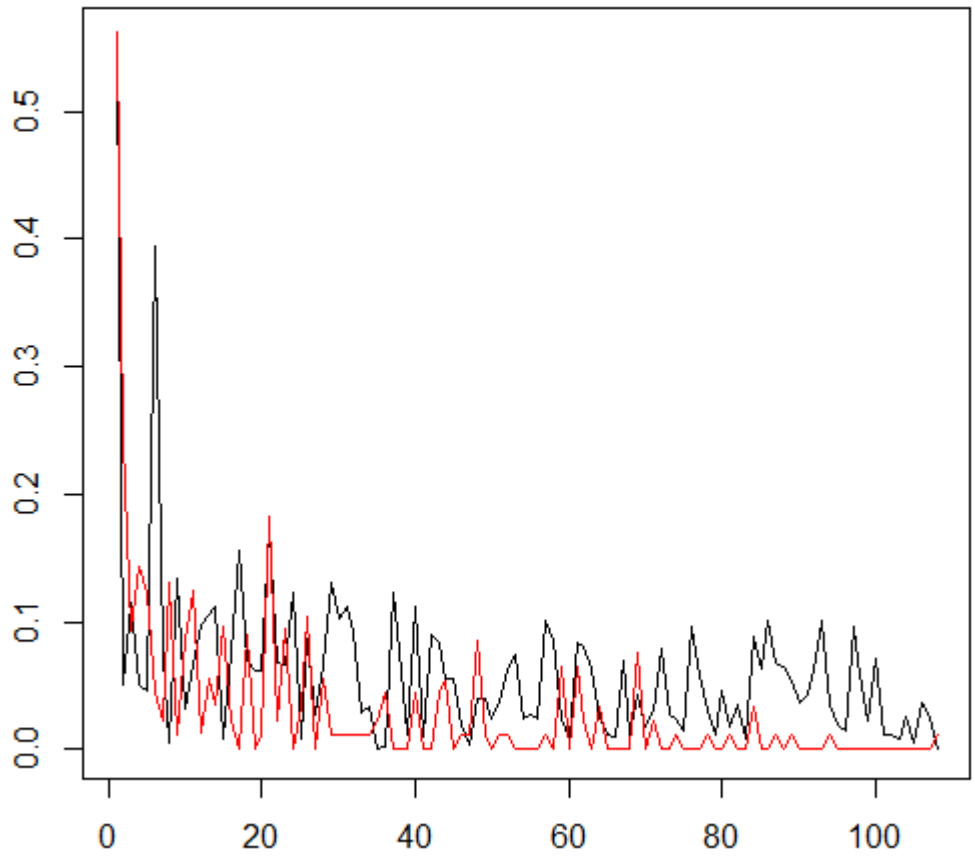
```
[109,] 0.57156795 0.5716053 0.560340791

[28,] 0.59897065 0.5716053 0.571605347
```

Διαισθητικά φαίνεται ότι η μέθοδος της πιθανοφάνειας βρήκε πιο γρήγορα επιθυμητό δείγμα. Το παρακάτω γράφημα απεικονίζει τον πίνακα M όπου με **κόκκινο** χρώμα είναι το **CV**, με **μαύρο** το **CVdeig.**, και με **πράσινο** το **MLE(CV)** , όπου με ***** έχουμε την πρώτη φορά που πετυχαίνουμε επιθυμητή ακρίβεια με το **CVdeig.**, και με ***** την πρώτη φορά που πετυχαίνουμε επιθυμητή ακρίβεια με **MLE(CV) .**



Γράφημα 4 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Ομοιόμορφη διακριτή U(1,100)

Παρατηρούμε ότι η μέθοδος της πιθανοφάνειας έδωσε καλύτερα αποτελέσματα και φαίνεται να συγκλίνει. Αυτό φαίνεται και παρακάτω με το γράφημα των απόλυτων σφαλμάτων όπου με **κόκκινο** χρώμα είναι τα απόλυτα σφάλματα με τη μέθοδο της **πιθανοφάνειας** και με **μαύρο χωρίς**.



Γράφημα 5: Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Ομοιόμορφη διακριτή U(1,100)

Όλα τα παραπάνω «συμπυκνώθηκαν» στη συνάρτηση

ff (Par, Dist,ERROR,REP,PLOT)

Περιγραφή της ff (Par, Dist,ERROR,REP,PLOT)

Η συνάρτηση ff δέχεται τις παραμέτρους Par, Dist,ERROR,REP,PLOT όπου :

Par : Οι παράμετροι της κατανομής

Dist : η κατανομή

ERROR : το επιθυμητό απόλυτο σφάλμα

REP : επιλογή 1 αν έχω δειγματοληψία με επανάθεση και 0 αν δεν έχω

PLOT : επιλογή 1 αν θέλω γραφήματα και 0 αν δεν θέλω γραφήματα.

Σαν αποτελέσματα η ff δίνει :

1) Τον αριθμό του δείγματος που πρέπει να επιλεγεί έτσι ώστε |CVdeig.- CV |< ERROR.

2) Τον αριθμό του δείγματος που πρέπει να επιλεγεί έτσι ώστε |MLE(CV)- CV |< ERROR

3) Το MSE(CV) ,δηλαδή το μέσο τετραγωνικό σφάλμα της εκτίμησης με CVdeig.

4) Το MSE(MLE(CV)) ,δηλαδή το μέσο τετραγωνικό σφάλμα της εκτίμησης με MLE(CV).

Στο παράδειγμα μας

```
> ff(c(1,100),"Uniform",0.0001,1,0)
n me deigmatiko CV       n me MLE(CV)           MSE(CV)       MSE(MLE(CV))
       93.00000000        13.00000000        0.01582309         0.00551405
```

Πίνακας 6 :Αποτελέσματα της ff για Ομοιόμορφη διακριτή U(1,100) με επανάθεση

Παρατηρούμε ότι NCV=93 ενώ NMLE(CV)=13 και MSE(CV)=0.0158, MSE(MLE(CV))=0.0055 αποτελέσματα που δείχνουν ότι ο εκτιμητής MLE(CV) είναι καλύτερος από τον CVdeig.

Ωστόσο κάθε φορά που τρέχουμε τον αλγόριθμο δε παίρνουμε πάντα το ίδιο αποτέλεσμα(βλέπε NCV=128 και 93,NMLE(CV)=28 και 13) αλλά παραπλήσιο. Για αυτό πρέπει να τρέξουμε τη διαδικασία πολλές φορές και για την ακρίβεια πάνω από 30(εμπειρικά) έτσι ώστε χρησιμοποιώντας το Κ.Ο.Θ. να βρούμε ένα διάστημα εμπιστοσύνης για το μέγεθος του δείγματος. Την διαδικασία υλοποιεί η συνάρτηση PROS(συνάρτηση προσομοίωσης).

Περιγραφή της PROS (N, Par, Dist, ERROR, REP)

Η PROS δέχεται τις παραμέτρους N, Par, Dist, ERROR, REP όπου :

N : το πλήθος των επαναλήψεων της διαδικασίας

Par : οι παράμετροι της κατανομής

Dist : η κατανομή

ERROR : το επιθυμητό απόλυτο σφάλμα

REP : επιλογή 1 αν έχω δειγματοληψία με επανάθεση και 0 αν δεν έχω .

Σαν αποτέλεσμα η συνάρτηση PROS μας δίνει

1) το 99% διάστημα εμπιστοσύνης του αριθμού του δείγματος που πρέπει να επιλεγεί έτσι ώστε |CVdeig.- CV |< ERROR.

2) το 99% διάστημα εμπιστοσύνης του αριθμού του δείγματος που πρέπει να επιλεγεί έτσι ώστε |MLE(CV)- CV |< ERROR

3) το 99% διάστημα εμπιστοσύνης για το MSE(CV) ,δηλαδή του μέσου τετραγωνικού σφάλματος της εκτίμησης με CVdeig.

4) το 99% διάστημα εμπιστοσύνης για το MSE(MLE(CV)) ,δηλαδή του μέσου τετραγωνικού σφάλματος της εκτίμησης με MLE(CV).

Στο παράδειγμα μας

```
> PROS(1000,c(1,100),"Uniform",0.0001,1)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              336.2708                377.7152

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        23.20298          24.72102

[[3]]
MIN MSE(CV) MAX MSE(CV)
0.004724860 0.005794194

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
     0.002909772       0.003678956
```

Πίνακας 7 : Αποτελέσματα της PROS για Ομοιόμορφη διακριτή U(1,100) με επανάθεση

Παρατηρούμε ότι το 99% δ.ε για το CVdeig.= (336.2708, 377.7152) για να έχω επιθυμητό αποτέλεσμα ενώ το αντίστοιχο 99% δ.ε για το MLE(CV) =(23.20298,24.72102) , κάτι που δείχνει ότι η εκτίμηση με το MLE(CV) είναι καλύτερη. Τα ίδια αποτελέσματα παίρνουμε και αν δούμε το 99% δ.ε. για το MSE(CV)=(0.00472486,0.005794194) και για το MSE(MLE(CV)) =(0.002909772,0.003678956) .

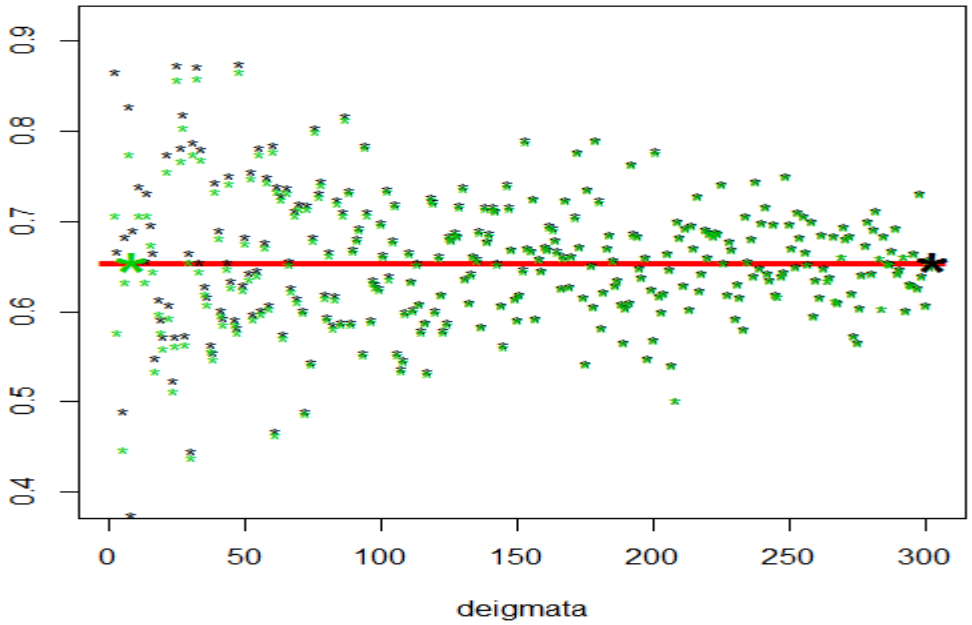
4.2 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ BERNOULLI ΚΑΤΑΝΟΜΗ

Συνεχίζουμε με την κατανομή Bernoulli. Θα τρέξουμε μόνο τις συναρτήσεις ff και PROS αφού περιγράψαμε αναλυτικά στην προηγούμενη ενότητα τις λεπτομέρειες του προγράμματος. Τρέχουμε την συνάρτηση ff και έχουμε:

```
> ff(0.7,"Bernoulli",0.0001,1,1)
n me deigmatiko CV       n me MLE(CV)            MSE(CV)       MSE(MLE(CV))
      3.040000e+02       1.000000e+01       9.572161e-03       7.590610e-03
```

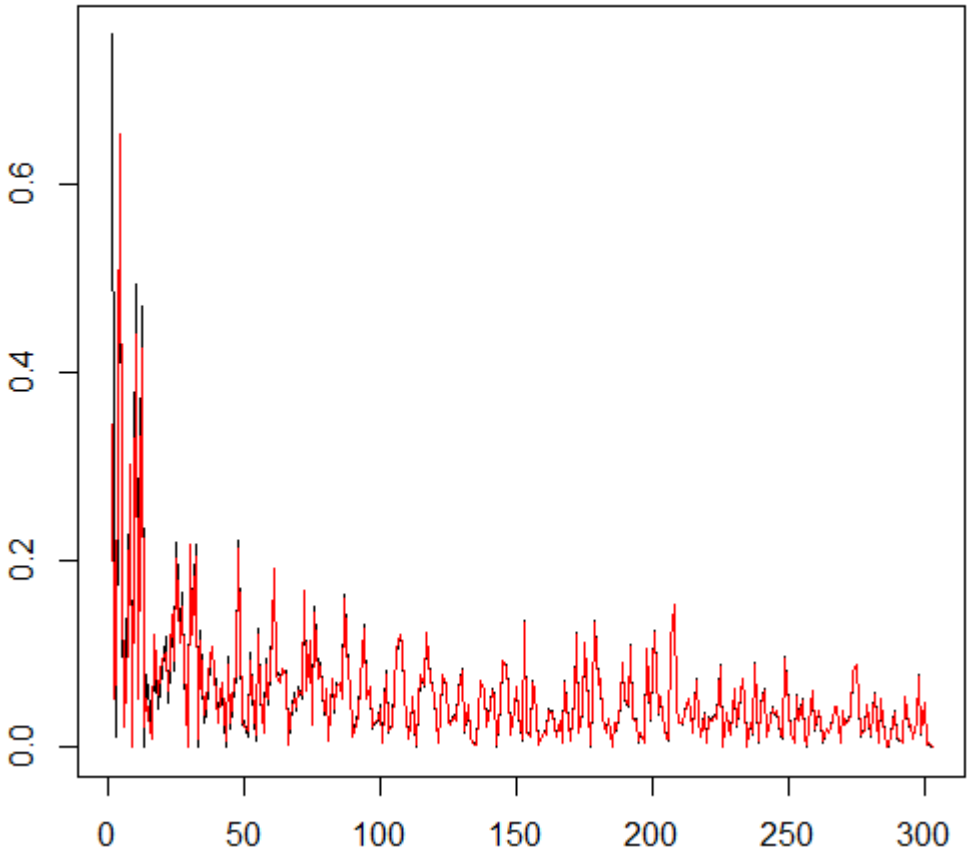
Πίνακας 8 : Αποτελέσματα της ff για Bernoulli B(1,0.7) με επανάθεση

Παρατηρούμε ότι η εκτίμηση με το MLE(CV) είναι καλύτερη με μεγάλη διαφορά στον αριθμό του δείγματος. Παρακάτω το γράφημα των συντελεστών μεταβλητότητας.



Γράφημα 6 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Bernoulli B(1,0.7)

Παρατηρούμε ότι το τα σημεία του *CVdeig.* είναι «πολύ κοντά» με αυτά του MLE(CV), «ελαφρώς» μετατοπισμένα, ωστόσο η επιλογή του κατάλληλου δείγματος γίνεται πάλι πιο γρήγορα με τη μέθοδο της Πιθανοφάνειας.



Γράφημα 7 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Bernoulli B(1,0.7)

Όπως περιμέναμε επειδή τα σημεία είναι «πολύ κοντά» μεταξύ τους δεν υπάρχει διαφορά στα σφάλματα. Επίσης παρατηρείται σύγκλιση.

Τρέχουμε την συνάρτηση PROS και έχουμε

```
> PROS(1000,0.7,"Bernoulli",0.0001,1)
```

```
[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
               345.3166                386.5594

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
         86.99502         117.01098

[[3]]
MIN MSE(CV) MAX MSE(CV)
   307.1639    509.3452

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
         307.1617          509.3432
```

Πίνακας 9:Αποτελέσματα της PROS για Bernoulli B(1,0.7) με επανάθεση

Παρατηρούμε ότι με τη μέθοδο της πιθανοφάνειας πρέπει να πάρουμε μικρότερο δείγμα. Το μέσο τετραγωνικό σφάλμα είναι σχεδόν ίδιο επειδή κάποιες φορές η μέση τιμή βρέθηκε 0, οπότε το CVdeig.→ ∞ , MLE(CV) → ∞ και για προγραμματιστικούς λόγους το θέσαμε 1000.

4.3 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Πριν προχωρήσουμε να αναφέρουμε ότι στη Διωνυμική κατανομή υπάρχει ο περιορισμός (στο πρόγραμμα) n ≤170 καθώς η R δεν μπορεί να υπολογίσει n! για n> 170. Συνεπώς επιλέγουμε το πολύ n=170

Τρέχουμε την ff και έχουμε

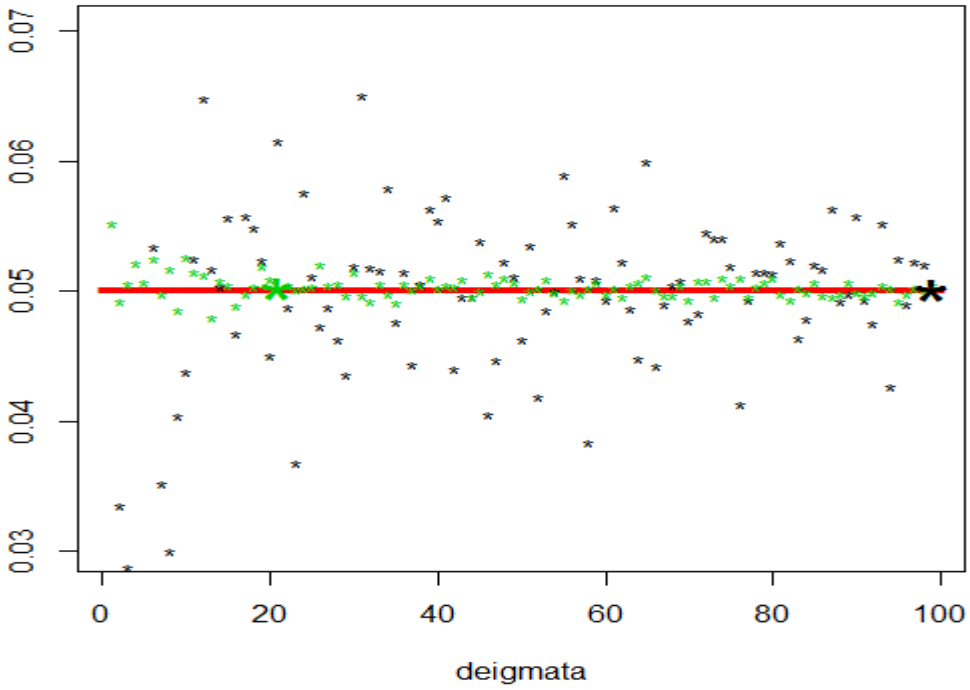
```
> ff(c(170,0.7),"Binomial",0.0001,1,1)
[1] 1.000000e+02 2.200000e+01 5.799587e-05 8.664544e-07
```

Πίνακας 10 : Αποτελέσματα της ff για Διωνυμική B(170,0.7) με επανάθεση

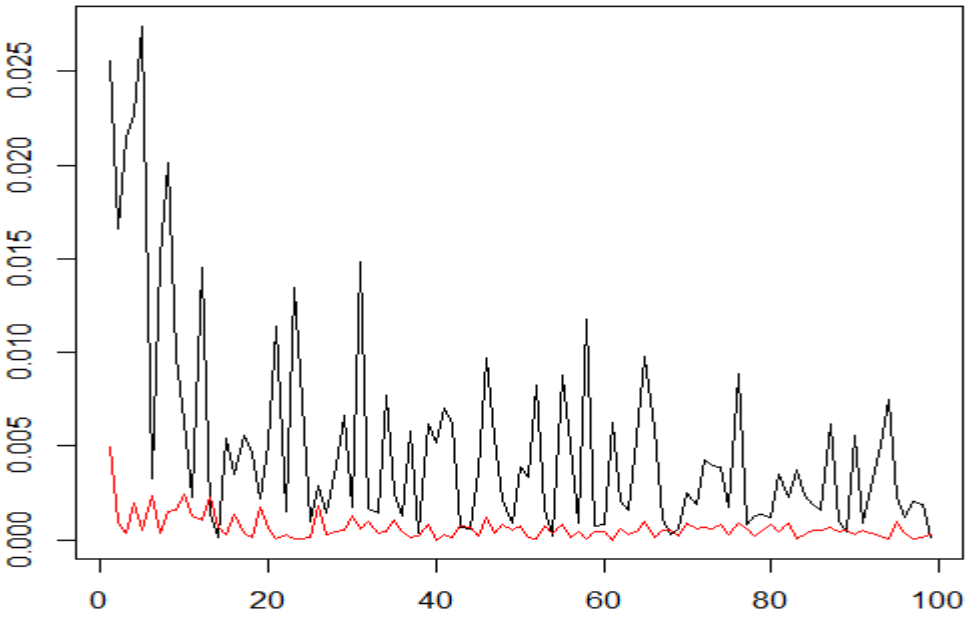
Πάλι παρατηρούμε ότι έχουμε καλύτερα αποτελέσματα με την μέθοδο της Πιθανοφάνειας διότι NCV=100 ,NMLE(CV)=22. Επίσης MSE(CV) = 5.799587 10^{-5} ενώ

MSE(MLE(CV))=8.664544 10^{-7}.

Η μέθοδος της Πιθανοφάνειας δηλαδή έχει μικρότερο μέσο τετραγωνικό σφάλμα και βρίσκει κατάλληλο αριθμό δείγματος πιο γρήγορα κάτι που φαίνεται και στα παρακάτω γραφήματα.



Γράφημα 8 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Διωνυμική B(170,0.7)



Γράφημα 9 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Διωνυμική B(170,0.7)

Εδώ τα απόλυτα σφάλματα με τη μέθοδο της Πιθανοφάνειας είναι πολύ μικρά και ίσως ο εκτιμητής MLE(CV) να είναι αμερόληπτος ή σχεδόν αμερόληπτος.

Τρέχουμε τη συνάρτηση PROS και έχουμε :

```
> PROS(1000,c(170,0.7),"Binomial",0.0001,1)

 [[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              68.28576                75.98824

 [[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
         15.99832          17.80768

 [[3]]
 MIN MSE(CV)   MAX MSE(CV)
0.0001014833 0.0001143447

 [[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    1.144287e-06     1.286262e-06
```

Πίνακας 11: Αποτελέσματα της PROS για Διωνυμική B(170,0.7) με επανάθεση

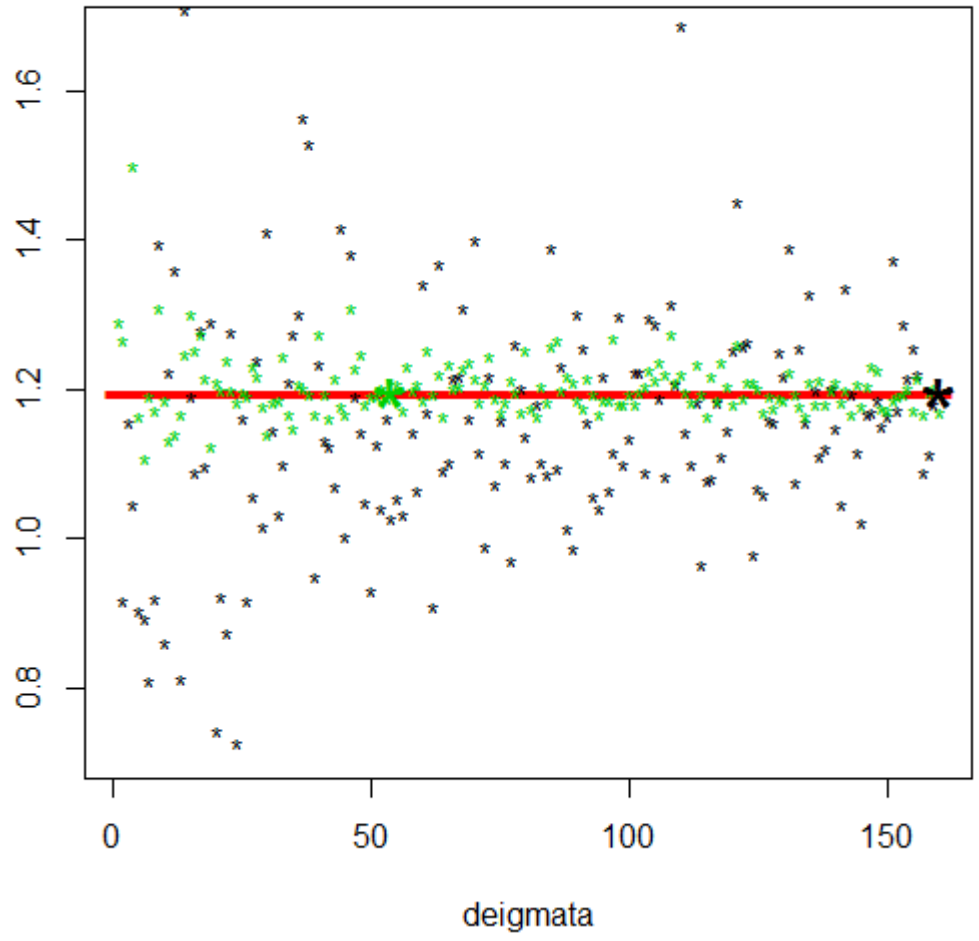
4.4 *ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ*

Τρέχουμε την ff και έχουμε

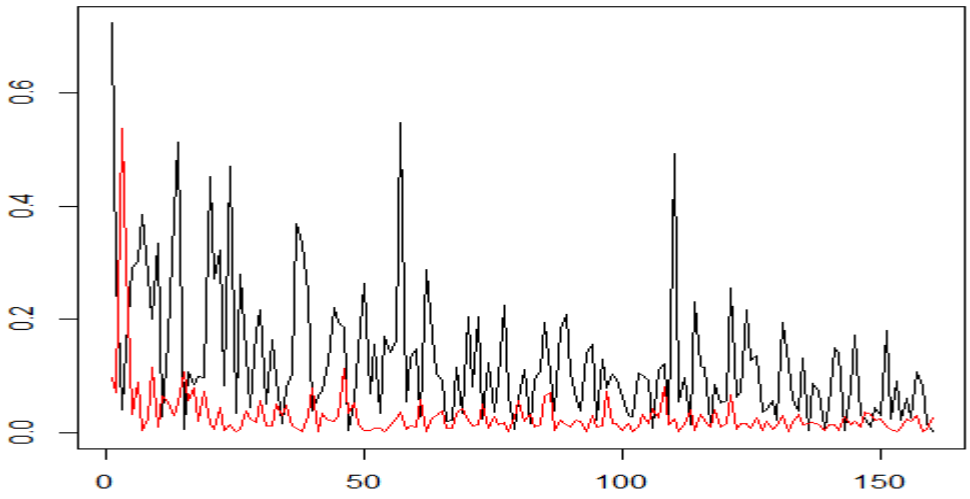
```
> ff(0.3,"Geometrical",0.001,1,1)
n me deigmatiko CV       n me MLE(CV)            MSE(CV)       MSE(MLE(CV))
      1.610000e+02       5.500000e+01       3.056122e-02       3.540949e-03
```

Πίνακας 12 : Αποτελέσματα της ff για Γεωμετρική με p=0.3 με επανάθεση

Πάλι καλύτερα αποτελέσματα με την μέθοδο της Πιθανοφάνειας καθώς χρειάζεται να επιλέξουμε μικρότερο δείγμα για να πετύχουμε επιθυμητή ακρίβεια. Επίσης το μέσο τετραγωνικό σφάλμα είναι πάλι μικρότερο. Παρακάτω τα γραφήματα που αποτυπώνουν την εικόνα αυτή.



Γράφημα 10 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Γεωμετρική με p=0.3



Γράφημα 11 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Γεωμετρική με p=0.3

Η σύγκλιση του MLE(CV) στο CV είναι αρκετά καλή και γρήγορη.

Τρέχουμε τη συνάρτηση PROS και έχουμε :

```
> PROS(1000,0.3,"Geometrical",0.001,1)

 [[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
               140.7348                156.8772

 [[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
         44.09711          49.76489

 [[3]]
MIN MSE(CV) MAX MSE(CV)
     918.836    2700.932

 [[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
          918.801          2700.897
```

Πίνακας 13: Αποτελέσματα της PROS για Γεωμετρική με p=0.3 με επανάθεση

Παρατηρούμε ότι θα χρειαστούμε πάλι με τη μέθοδο της Πιθανοφάνειας μικρότερο δείγμα για να πετύχουμε την επιθυμητή ακρίβεια. Επιπλέον τα μέσα τετραγωνικά σφάλματα είναι ίδια γιατί σε κάποια δείγματα είχαμε μέση τιμή 0.

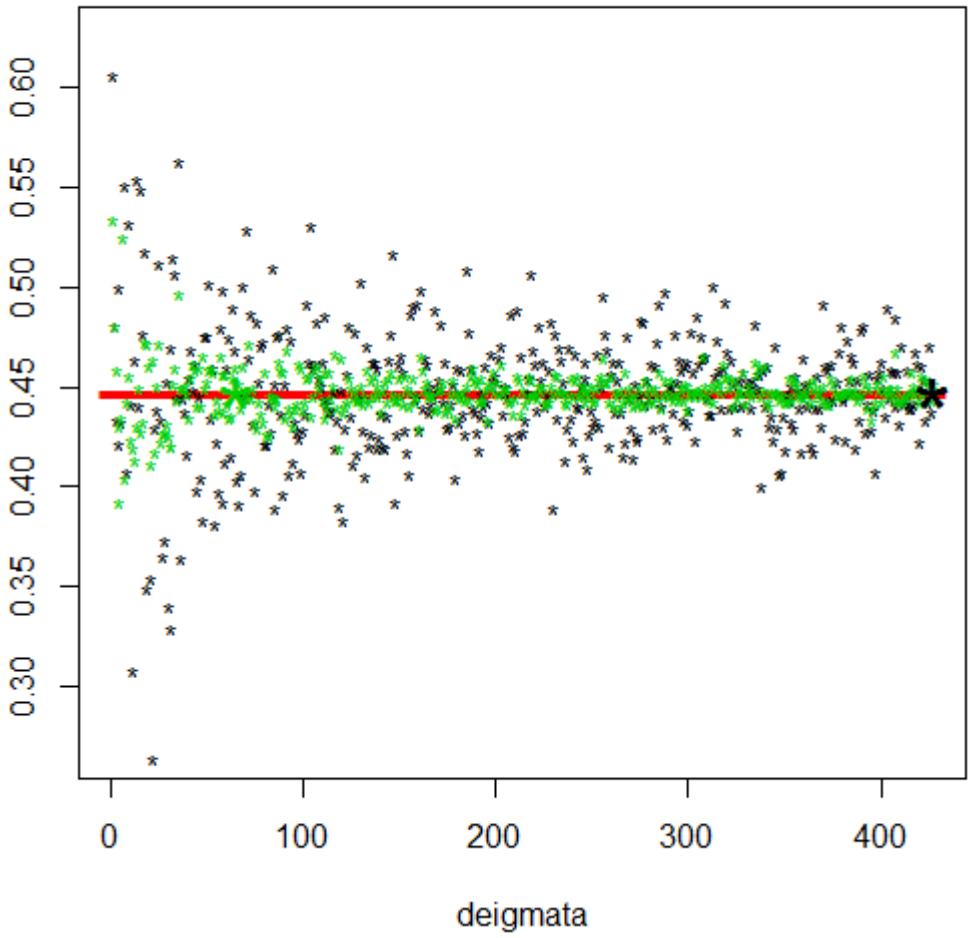
4.5 *ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ POISSON ΚΑΤΑΝΟΜΗ*

Τρέχουμε την ff και έχουμε

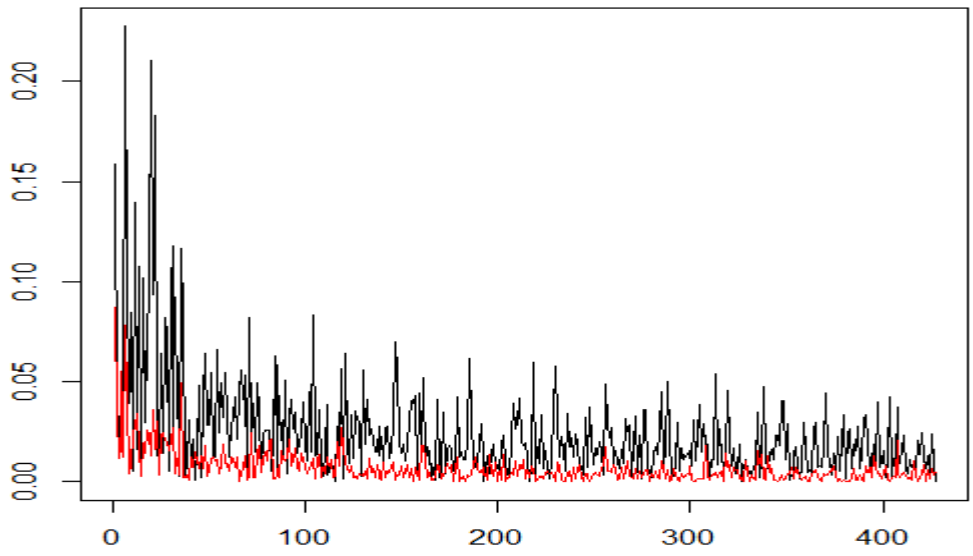
```
> ff(5,"Poisson",0.0001,1,1)
n me deigmatiko CV       n me MLE(CV)            MSE(CV)       MSE(MLE(CV))
     4.280000e+02       6.700000e+01       1.322448e-03       1.288502e-04
```

Πίνακας 14 : Αποτελέσματα της ff για Poisson, λ=5 με επανάθεση

Πάλι καλύτερα αποτελέσματα με την Πιθανοφάνεια κάτι που φαίνεται και από τα παρακάτω γραφήματα



Γράφημα 12 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Poisson με λ=5



Φαίνεται ξανά η συνέπεια του MLE(CV).

Γράφημα 13 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Poisson με λ=5

Τρέχουμε τη συνάρτηση PROS και έχουμε :

```
> PROS(1000,5,"Poisson",0.0001,1)

 [[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
               267.518                300.730

 [[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        19.83868         26.41132

 [[3]]
MIN MSE(CV) MAX MSE(CV)
0.003938976 0.004959834

 [[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    0.0003532278     0.0004615737
```

Πίνακας 15: Αποτελέσματα της PROS για Poisson, λ=5 με επανάθεση.

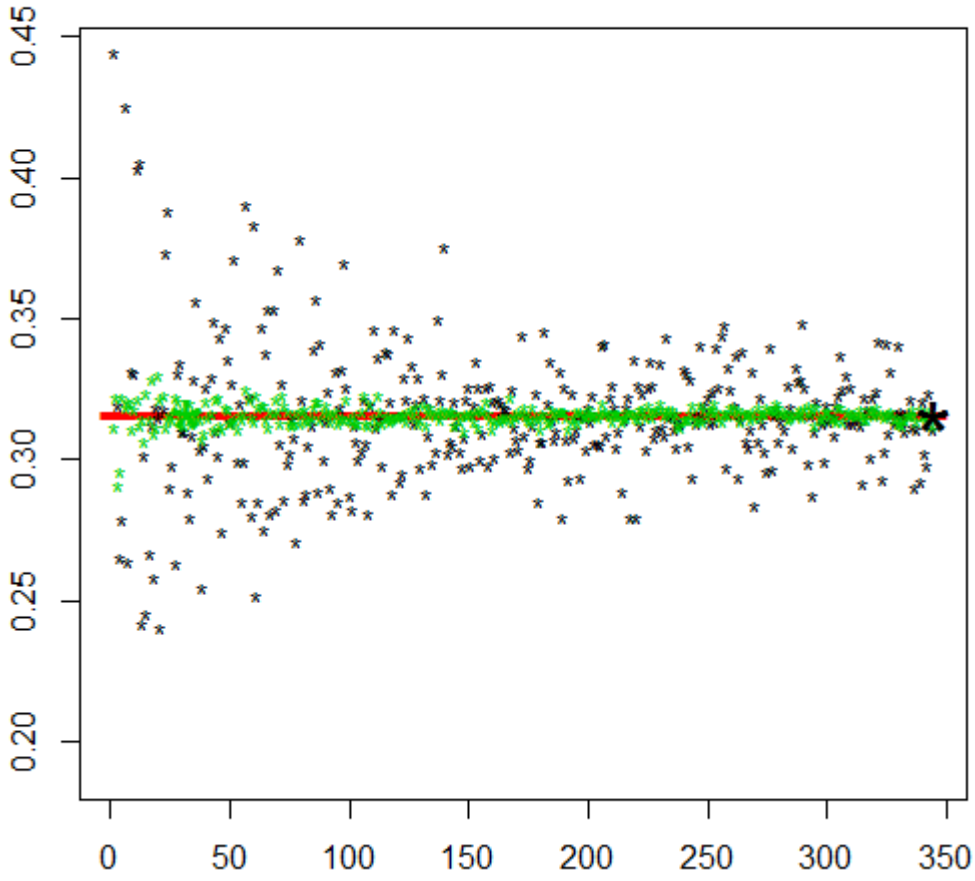
Η μέθοδος της Πιθανοφάνεια είναι ξανά αποδοτικότερη.

4.6 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΑΡΝΗΤΙΚΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

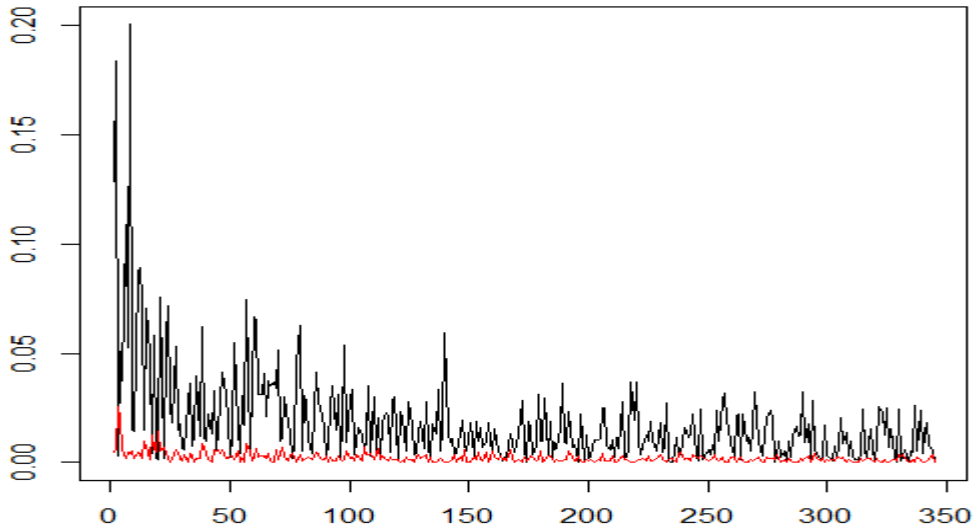
Τρέχουμε την ff και έχουμε

```
> ff(c(25,0.6),"NBinomial",0.0001,1,1)
n me deigmatiko CV       n me MLE(CV)            MSE(CV)       MSE(MLE(CV))
      3.460000e+02       3.300000e+01       8.168072e-04       1.156664e-05
```

Πίνακας 16 : Αποτελέσματα της ff για Αρνητική Διωνυμική με ν=25, p=0.6 με επανάθεση



Γράφημα 14 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Αρνητική Διωνυμική με ν=25 και p=0.6



Γράφημα 15 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Αρνητική Διωνυμική με ν=25 , p=0.6

Ο εκτιμητής MLE(CV) μοιάζει αμερόληπτος ή σχεδόν αμερόληπτος.

Τρέχουμε τη συνάρτηση PROS και έχουμε :

```
> PROS(1000,c(25,0.6),"NBinomial",0.0001,1)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
               236.0126                262.7654

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
         55.01784          61.51616

[[3]]
MIN MSE(CV) MAX MSE(CV)
0.001561708 0.001791888

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    2.534337e-05     2.928410e-05
```

Πίνακας 17: Αποτελέσματα της PROS για Αρνητική Διωνυμική με ν=25, p=0.6 με επανάθεση.

Αρκετά «καλή» οικονομία δείγματος με NMLE(CV) ≅ 58 αντί για NCV≅ 250.

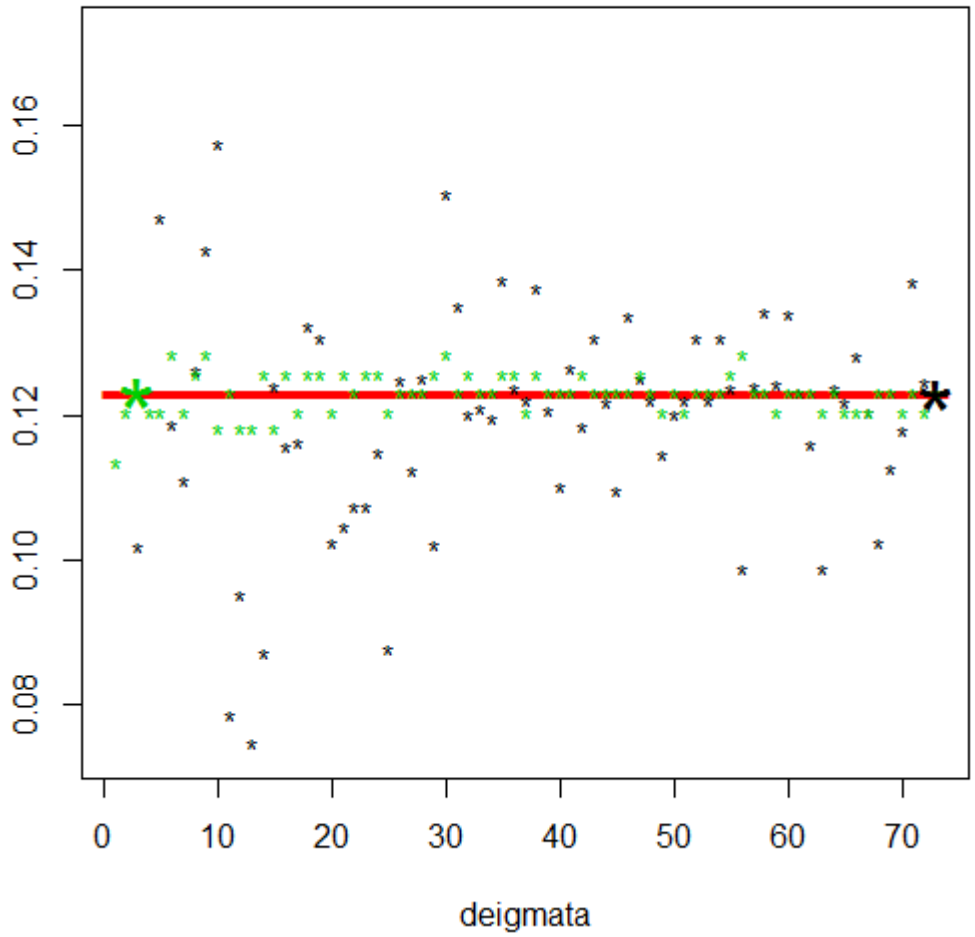
4.7 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ.

Τρέχουμε την ff και έχουμε

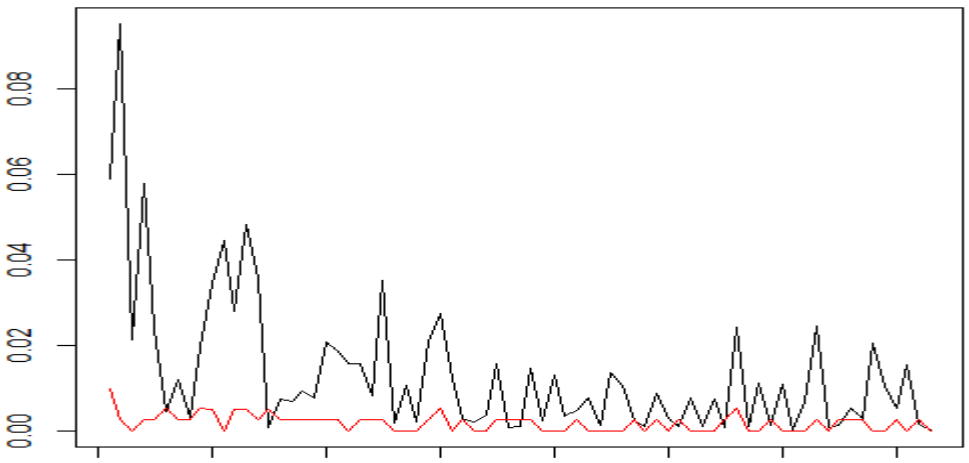
```
> ff(c(100,40,50),"HGeometrical",0.0001,1,1)
[1] 7.400000e+01 4.000000e+00 4.498512e-04 7.156316e-06
```

Πίνακας 18 : Αποτελέσματα της ff για Υπεργεωμετρική με N=100, K=40, ν=50 με επανάθεση

Και σε αυτή τη περίπτωση ο MLE(CV) είναι καλύτερος εκτιμητής από το CVdeig. και παρατηρείται σύγκλιση.



Γράφημα 16 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Υπεργεωμετρική με N=100, K=40, v=50



Γράφημα 17 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Υπεργεωμετρική με N=100, K=40, ν=50 με επανάθεση

Εδώ επιβεβαιώνεται η θεωρία ότι δηλαδή π ο MLE(CV) είναι αμερόληπτος εκτιμητής.

Τρέχουμε τη συνάρτηση PROS και έχουμε :

```
> PROS(1000,c(100,40,50),"HGeometrical",0.0001,1)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              124.0999                138.6361

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        6.640265         7.315735

[[3]]
 MIN MSE(CV)  MAX MSE(CV)
0.0004521588 0.0005425371

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    8.594551e-06     1.044220e-05
```

Πίνακας 19: Αποτελέσματα της PROS για Υπεργεωμετρική με N=100, K=40, ν=50 με επανάθεση.

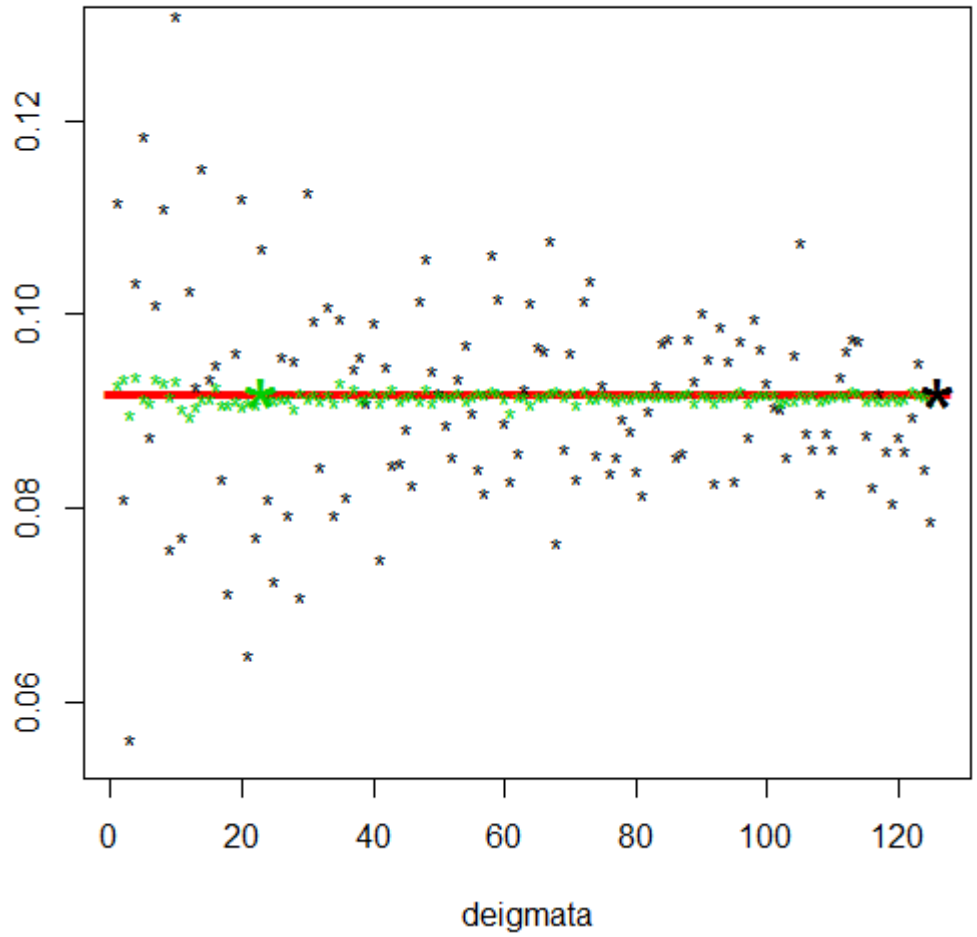
Εμφανίζονται ξανά οι γνωστές ιδιότητες των Ε.Μ.Π.

4.8 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΜΕ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΑΡΝΗΤΙΚΗ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

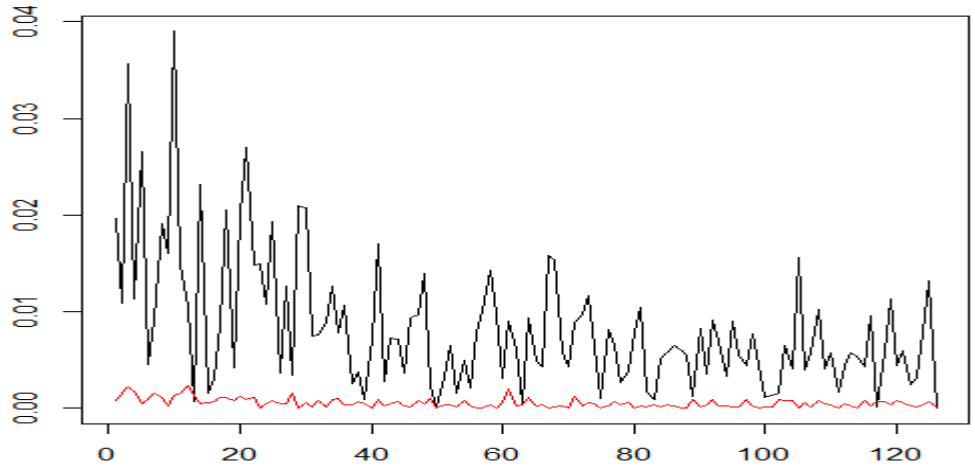
Τρέχουμε την ff και έχουμε :

```
> ff(c(20,30,60),"NHGeometrical",0.0001,1,1)
n me deigmatiko CV       n me MLE(CV)            MSE(CV)       MSE(MLE(CV))
     1.270000e+02       2.400000e+01       1.155486e-04       5.443291e-07
```

Πίνακας 20 : Αποτελέσματα της ff για Αρνητική Υπεργεωμετρική με ν=20, K=30, N=60 με επανάθεση



Γράφημα 18 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα με επανάθεση στην Αρνητική Υπεργεωμετρική με ν=20, K=30, N=60



Γράφημα 19 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία με επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Αρνητική Υπεργεωμετρική με ν=20, K=30, N=60

Και εδώ επιβεβαιώνεται η θεωρία ότι δηλαδή ο MLE(CV) είναι αμερόληπτος εκτιμητής

Τρέχουμε τη συνάρτηση PROS και έχουμε :

```
> PROS(1000,c(20,30,60),"NHGeometrical",0.0001,1)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              96.58937               108.05663

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        13.70154         15.42846

[[3]]
 MIN MSE(CV)   MAX MSE(CV)
0.0002812887 0.0003299635

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    9.090709e-07     1.045090e-06
```

Πίνακας 21: Αποτελέσματα της PROS για Αρνητική Υπεργεωμετρική με ν=20, K=30, N=60, με επανάθεση

Μεγάλη «οικονομία» στη δειγματοληψία και σε αυτή τη περίπτωση.

ΑΝΑΚΕΦΑΛΑΙΩΣΗ

Συνοψίζοντας στο κεφάλαιο 4 θεμελιώσαμε στο προγραμματιστικό περιβάλλον της R, την θεωρία των προηγούμενων κεφαλαίων στη περίπτωση της δειγματοληψίας με επανάθεση. Επιβεβαιώσαμε ότι ο MLE(CV) είναι καλύτερος εκτιμητής από τον CVdeig. και επίσης παρατηρήσαμε ότι έχει όλες εκείνες τις ασυμπτωτικές ιδιότητες των Ε.Μ.Π., δηλαδή της συνέπειας της ασυμπτωτικής αμεροληψίας και της ασυμπτωτικής αποδοτικότητας.Ο MLE(CV) σε όλες τις περιπτώσεις είχε μικρότερο μέσο τετραγωνικό σφάλμα και βρέθηκε κατάλληλο δείγμα πολύ πιο γρήγορα.

ΚΕΦΑΛΑΙΟ 5 «ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΣΤΟ ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ ΤΗΣ R»

ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο 5 θα προσομοιώσουμε την θεωρία που αναπτύχθηκε στα προηγούμενα κεφάλαια ομοίως όπως στο κεφάλαιο 4.Θα επικεντρωθούμε στη δειγματοληψία χωρίς επανάθεση για τις ίδιες διακριτές κατανομές που μελετήθηκαν στο προηγούμενο κεφάλαιο και στην επιλογή του κατάλληλου μεγέθους δείγματος έτσι ώστε να έχουμε επιθυμητό απόλυτο σφάλμα προσέγγισης.

Θα επιλέγεται πάλι ένα απόλυτο σφάλμα προσέγγισης του CV(του πληθυσμού) από το CVdeig.(δειγματικό) και από το MLE(CV) (δειγματικό με τη μέθοδο της πιθανοφάνειας). Έπειτα με υπολογιστικές μεθόδους ο αλγόριθμος θα δίνει απάντηση για το τι δείγμα πρέπει να πάρουμε σε μορφή διαστήματος εμπιστοσύνης και για τις δύο περιπτώσεις.

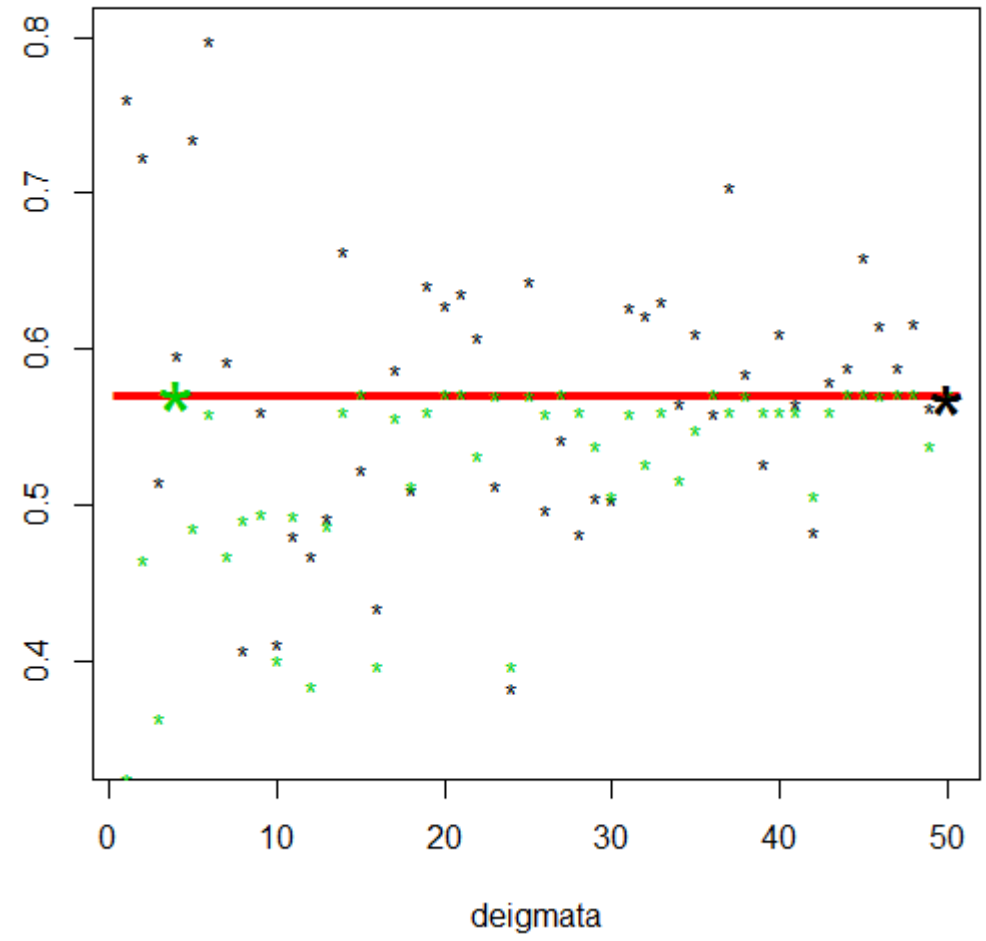
Πιο συγκεκριμένα ο αλγόριθμος σε κάθε βήμα παράγει δείγματα χωρίς επανάθεση με αυξανόμενο πλήθος (ξεκινώντας από 1) και υπολογίζει CV(που είναι σταθερό) , CV deig. και MLE(CV) .Έπειτα υπολογίζει τα απόλυτα σφάλματα και σταματάει όταν το απόλυτο σφάλμα γίνει για $1^{η}$ φορά. μικρότερο από αυτό που επέλεξε ο χρήστης. Η διαφορά με την μέθοδο με επανάθεση είναι ότι τώρα το μέγεθος του δείγματος που επιλέγεται δε μπορεί να είναι μεγαλύτερο από τον πληθάριθμο του συνόλου αναφοράς. Ως εκ τούτου μπορεί να μη βρεθεί κατάλληλο N, οπότε ο χρήστης πρέπει να αυξήσει το σφάλμα. Στα παρακάτω παραδείγματα έχει επιλεγεί κατάλληλο σφάλμα και έχουμε επιλέξει τις ίδιες παραμέτρους με το κεφάλαιο 4 για να εντοπίσουμε τυχόν διαφορές.

5.1 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΟΜΟΙΟΜΟΡΦΗ ΚΑΤΑΝΟΜΗ

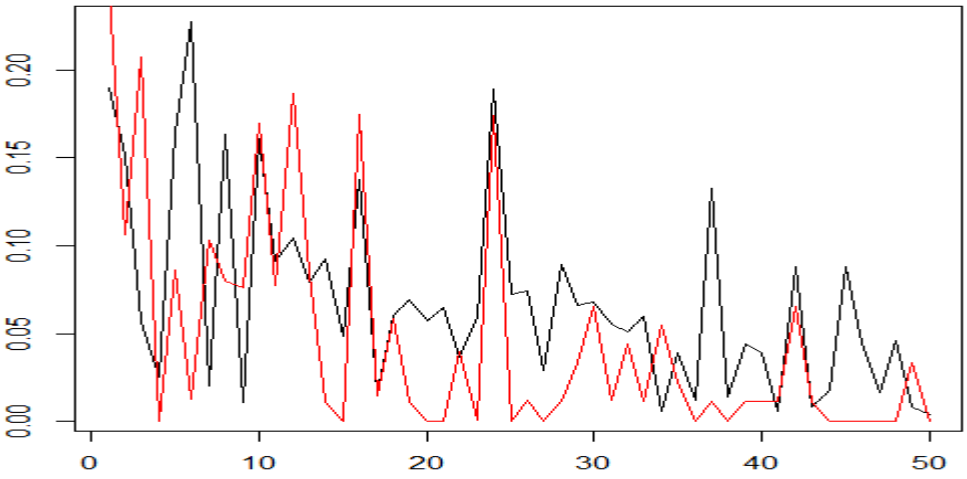
Τρέχουμε την ff όπου έχουμε στη θέση REP=0 (χωρίς επανάθεση) και έχουμε

```
> ff(c(1,100),"Uniform",0.006,0,1)
n me deigmatiko CV       n me MLE(CV)            MSE(CV)       MSE(MLE(CV))
      51.000000000        5.000000000       0.007747696        0.006115314
```

Πίνακας 22 : Αποτελέσματα της ff για Ομοιόμορφη διακριτή U(1,100) χωρίς επανάθεση.



Γράφημα 20 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Ομοιόμορφη διακριτή U(1,100)



Γράφημα 21 : Απόλυτα σφάλματα εκτίμησης CV για δειγματοληψία χωρίς επανάθεση, με ή χωρίς την μέθοδο της πιθανοφάνειας στην Ομοιόμορφη διακριτή U(1,100)

Η μέθοδος της Πιθανοφάνειας μοιάζει ελαφρώς καλύτερη.

Τρέχουμε τη συνάρτηση PROS με απόλυτο σφάλμα 0.006 και έχουμε :

```
> PROS(1000,c(1,100),"Uniform",0.006,0)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              20.84921                22.97079

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        12.04036          13.00764

[[3]]
MIN MSE(CV) MAX MSE(CV)
 0.02961633  0.03310645

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
      0.02616970       0.02882808
```

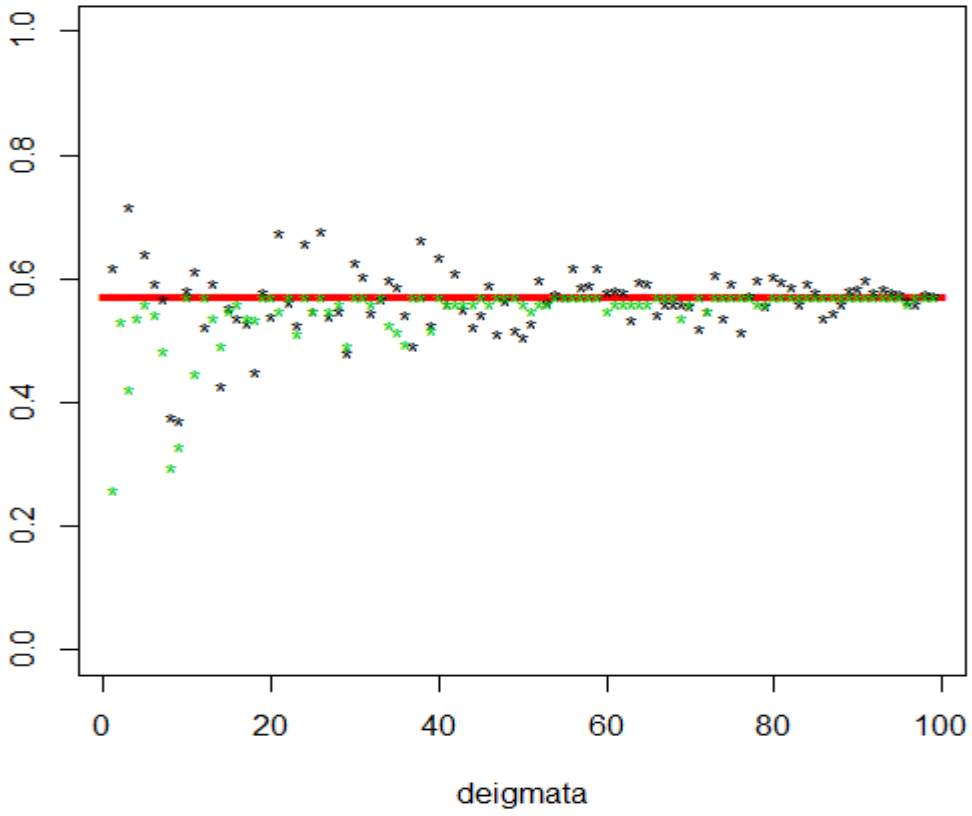
Πίνακας 23: Αποτελέσματα της PROS για Ομοιόμορφη διακριτή U(1,100) χωρίς επανάθεση

Παρατηρούμε ότι ενώ η μέθοδος της Πιθανοφάνειας δίνει καλύτερα αποτελέσματα το φαινόμενο δεν είναι το ίδιο «έντονο» στη δειγματοληψία χωρίς επανάθεση.

Ωστόσο παρατηρούμε ότι οι τιμές του NCV και του NMLE(CV) είναι «πολύ μικρές» κάτι που δείχνει ότι πολύ γρήγορα ικανοποιούνται τα κριτήρια τερματισμού.

|CVdeig.- CV |< ERROR και |MLE(CV)- CV |< ERROR.

Συνεπώς αφήνουμε ελεύθερο το γράφημα των CVdeig, MLE(CV) και έχουμε :



Γράφημα 22 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Ομοιόμορφη διακριτή U(1,100) ,χωρίς κριτήριο τερματισμού

Παρατηρούμε ότι πάλι η εκτίμηση με τη μέθοδο της Πιθανοφάνειας συγκλίνει.

5.2 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΚΑΤΑΝΟΜΗ BERNOULLI B(1, p)

Εδώ η διαδικασία δεν έχει νόημα καθώς Ω={0,1}.

Οπότε για τον υπολογισμό του CVdeig. θα επιλεγούν $x_1 = 0$, $x_2=1$ (αφού για n=1 δεν ορίζεται η δειγματική τυπική απόκλιση) και τότε $m=\frac{1}{2}$ και $s^2 = 0.5$ και

CVdeig. $= \frac{\sqrt{0.5}}{0.5} = \sqrt{2}$ και MLE(CV)$= \sqrt{\frac{1}{m} - 1} = 1$ και CV$= \sqrt{\frac{1}{P} - 1} > 0$

Προφανώς δε θα έχουμε καλή εκτίμηση σχεδόν σε όλες τις περιπτώσεις εκτός από $p=\frac{1}{3}$ για CVdeig. και p=0.5 για MLE(CV) με βέλτιστο μέγεθος δείγματος N=2.

Ακόμη για N=1 ορίζεται ο MLE(CV) αλλά μόνο αν επιλεγεί η τιμή $x_2=1$, οπότε MLE(CV)=0 και δεν είναι ποτέ καλός εκτιμητής.

5.3 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ B(n,p)

Εδώ θα τρέξουμε κατευθείαν τη συνάρτηση PROS καθώς έχουμε πολύ λίγα σημεία στα γραφήματα της ff, με απόλυτο σφάλμα 0.003.

```
PROS(30,c(170,0.7),"Binomial",0.003,0)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
               5.086687                7.646647

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
         2.217468          2.982532

[[3]]
 MIN MSE(CV)   MAX MSE(CV)
0.0002887104 0.0005060388

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
     3.668581e-06      8.326378e-06
```

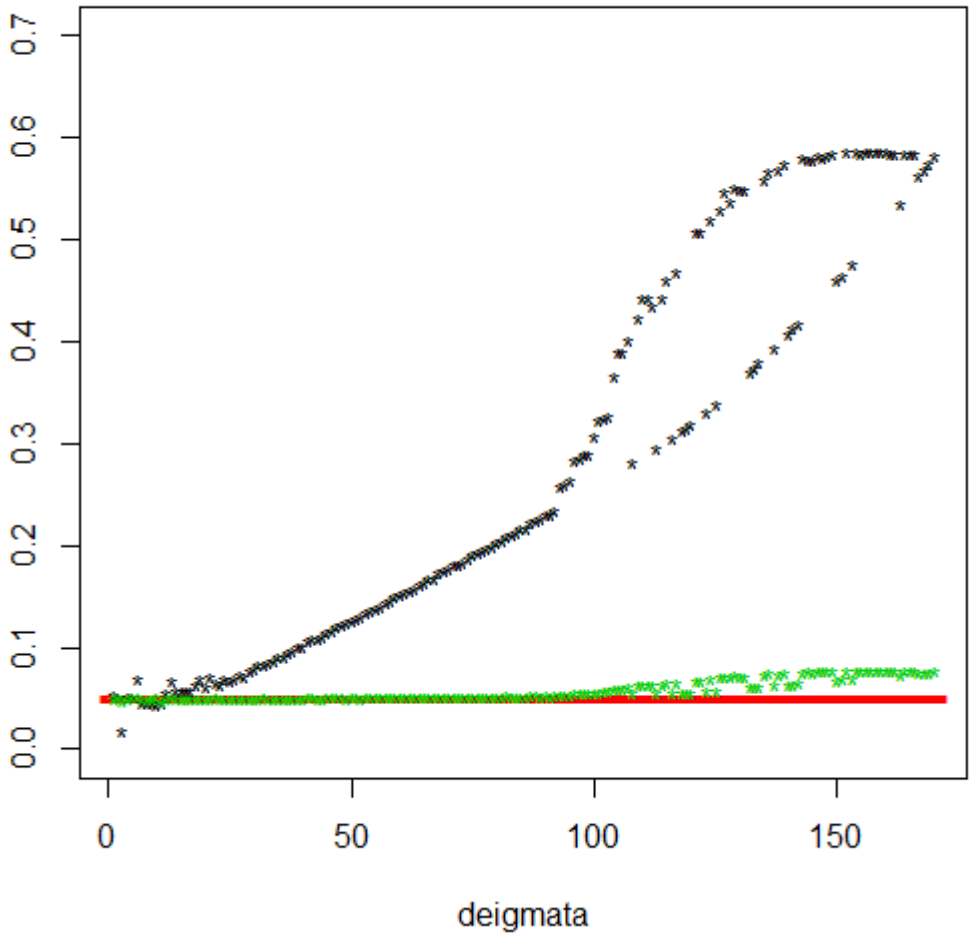
Πίνακας 24: Αποτελέσματα της PROS για Διωνυμική B(170,0.7) χωρίς επανάθεση

Πάλι έχουμε καλύτερα αποτελέσματα με τη μέθοδο της Πιθανοφάνειας αλλά το φαινόμενο δεν είναι το «έντονο» όσο στη περίπτωση με επανάθεση.

Ωστόσο παρατηρούμε ότι οι τιμές του NCV και του NMLE(CV) είναι «πολύ μικρές» κάτι που δείχνει ότι πολύ γρήγορα ικανοποιούνται τα κριτήρια τερματισμού.

|CVdeig.- CV |< ERROR και |MLE(CV)- CV |< ERROR.

Συνεπώς αφήνουμε ελεύθερο το γράφημα των CVdeig, MLE(CV) και βλέπουμε ότι



Γράφημα 23 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Διωνυμική B(170,0.7) ,χωρίς κριτήριο τερματισμού

Παρατηρούμε ότι καθώς μεγαλώνει το δείγμα έχουμε απόκλιση και με τις δύο μεθόδους, ενώ για πολύ μικρά δείγματα έχουμε σύγκλιση. Αυτό συμβαίνει γιατί καθώς αυξάνεται το δείγμα σε δειγματοληψία χωρίς επανάθεση τα «βάρη» (σχετικές συχνότητες) των στοιχείων μοιράζονται ομοιόμορφα στο δειγματοχώρο κάτι που δε συμβαίνει σε όλες τις κατανομές. Αξιοσημείωτο επίσης ότι για N<80 η μέθοδος της Πιθανοφάνειας δίνει καλά αποτελέσματα.

5.4 *ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ*

Θα τρέξουμε κατευθείαν τη συνάρτηση PROS καθώς έχουμε πολύ λίγα σημεία στα γραφήματα της ff. Επιλέγουμε αρκετά μεγάλο απόλυτο σφάλμα 0.25 καθώς δε βρίσκεται επιθυμητό N (και για τις 30 φορές προσομοίωσης) για μικρότερες τιμές του σφάλματος.

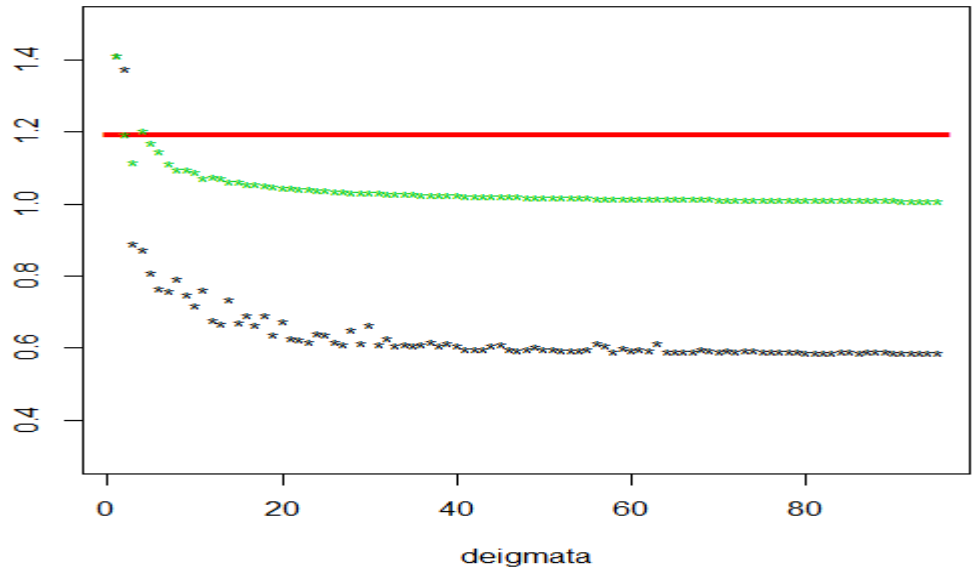
```
> PROS(30,0.3,"Geometrical",0.25,0)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
               2.156297                2.843703

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
          1.95655           2.24345

[[3]]
MIN MSE(CV) MAX MSE(CV)
 0.05321283  0.17641037

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
     0.005360955       0.048154869
```

Πίνακας 25: Αποτελέσματα της PROS για Γεωμετρική με p=0.3 χωρίς επανάθεση

Έχουμε καλύτερα αποτελέσματα με τη μέθοδο της Πιθανοφάνειας αλλά το φαινόμενο δεν είναι το «έντονο» όσο στη περίπτωση με επανάθεση καθώς εδώ δεν έχουμε μεγάλη ποσοστιαία διαφορά ανάμεσα στις δύο μεθόδους.

Ωστόσο παρατηρούμε ότι οι τιμές του NCV και του NMLE(CV) είναι «πολύ μικρές» κάτι που δείχνει ότι πολύ γρήγορα ικανοποιούνται τα κριτήρια τερματισμού.

Συνεπώς αφήνουμε ελεύθερο το γράφημα των CVdeig, MLE(CV) και βλέπουμε ότι



Γράφημα 24 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Γεωμετρική με p=0.3 ,χωρίς κριτήριο τερματισμού

Πάλι έχουμε απόκλιση καθώς αυξάνει η δειγματοληψία και προσέγγιση μόνο για πολύ μικρές τιμές δειγμάτων και μόνο για την μέθοδο της Πιθανοφάνειας.

5.5 *ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΚΑΤΑΝΟΜΗ POISSON P(λ).*

Πάλι θα τρέξουμε κατευθείαν τη συνάρτηση PROS καθώς έχουμε πολύ λίγα σημεία στα γραφήματα της ff. Επιλέγουμε μεγάλο απόλυτο σφάλμα 0.07 καθώς δε βρίσκεται επιθυμητό N (και για τις 30 φορές προσομοίωσης) για μικρότερες τιμές του σφάλματος.

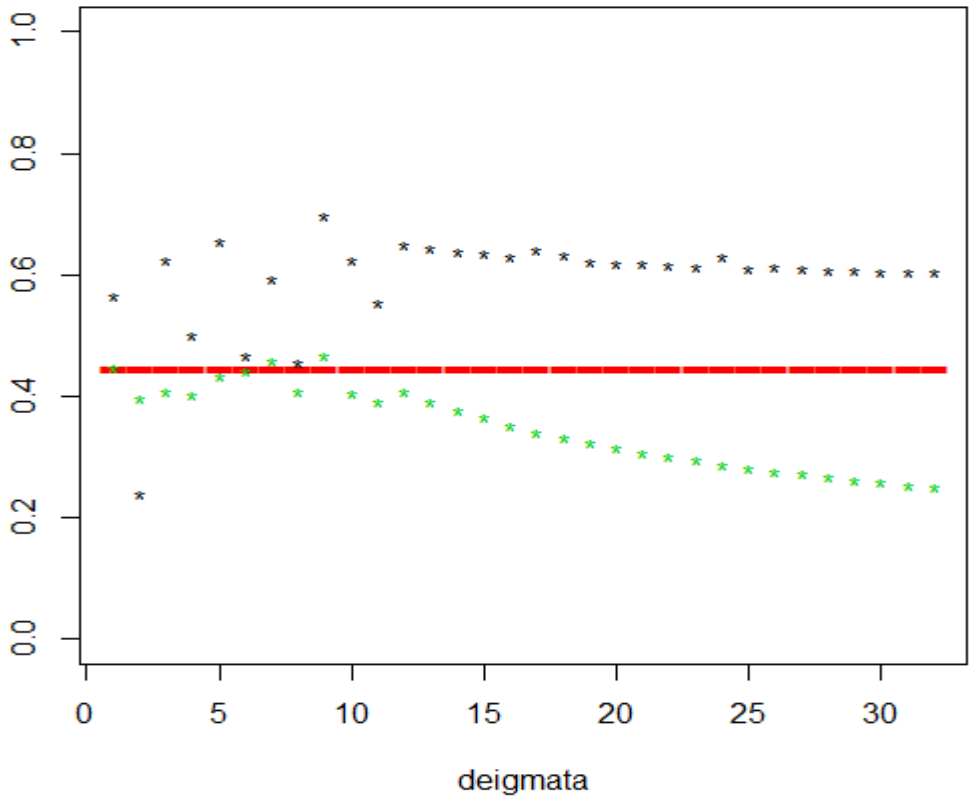
```
> PROS(30,5,"Poisson",0.07,0)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              3.345753                5.120914

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        2.031092          2.435575

[[3]]
MIN MSE(CV) MAX MSE(CV)
 0.01511448  0.05838127

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
     0.001632610       0.003797582
```

Πίνακας 26: Αποτελέσματα της PROS για Poisson με λ=5 χωρίς επανάθεση

Έχουμε καλύτερα αποτελέσματα με τη μέθοδο της Πιθανοφάνειας ωστόσο το φαινόμενο δεν είναι το «έντονο» όσο στη περίπτωση με επανάθεση.

Πάλι παρατηρούμε ότι οι τιμές του NCV και του NMLE(CV) είναι «πολύ μικρές» κάτι που δείχνει ότι πολύ γρήγορα ικανοποιούνται τα κριτήρια τερματισμού.

Συνεπώς αφήνουμε ελεύθερο το γράφημα των CVdeig, MLE(CV) και βλέπουμε ότι



Γράφημα 25 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Poisson με λ=5,χωρίς κριτήριο τερματισμού

Πάλι οι μέθοδοι εκτίμησης αποκλίνουν καθώς αυξάνει η δειγματοληψία με επανάθεση ενώ έχουμε καλή προσέγγιση για μικρό αριθμό δείγματος N<10 κυρίως με τη μέθοδο της Πιθανοφάνειας.

5.6 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΑΡΝΗΤΙΚΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Τρέχουμε τη συνάρτηση PROS .Επιλέγουμε απόλυτο σφάλμα 0.02 .

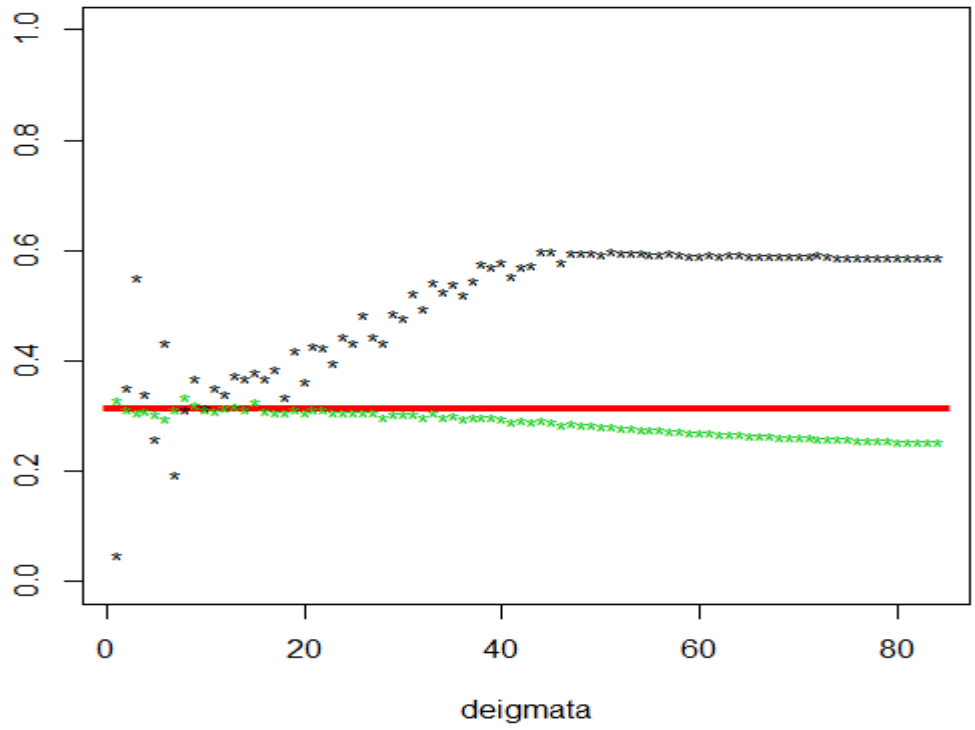
```
> PROS(30,c(25,0.6),"NBinomial",0.02,0)
```

```
[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
               4.929597                8.137070

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        2.215848          3.117485

[[3]]
MIN MSE(CV) MAX MSE(CV)
 0.01160664  0.02759778

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    0.0001923980     0.0003604452
```

Πίνακας 27: Αποτελέσματα της PROS για Αρνητική Διωνυμική με ν=25, p=0.6, χωρίς επανάθεση

Παρατηρούμε μικρές διαφορές στις δύο μεθόδους και ότι οι τιμές του NCV και του NMLE(CV) είναι περίπου 6 και 3 αντίστοιχα κάτι που δείχνει ότι πολύ γρήγορα ικανοποιούνται τα κριτήρια τερματισμού.

Συνεπώς αφήνουμε ελεύθερο το γράφημα των CVdeig, MLE(CV) και βλέπουμε ότι



Γράφημα 26 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Αρνητική Διωνυμική με ν=25, p=0.6, χωρίς επανάθεση, χωρίς κριτήριο τερματισμού

Παρατηρείται ξανά το φαινόμενο της απόκλισης, ωστόσο η μέθοδος της Πιθανοφάνειας είναι αρκετά «καλή» για n≤ 30.

5.7 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Τρέχουμε τη συνάρτηση PROS .Επιλέγουμε απόλυτο σφάλμα 0.02.

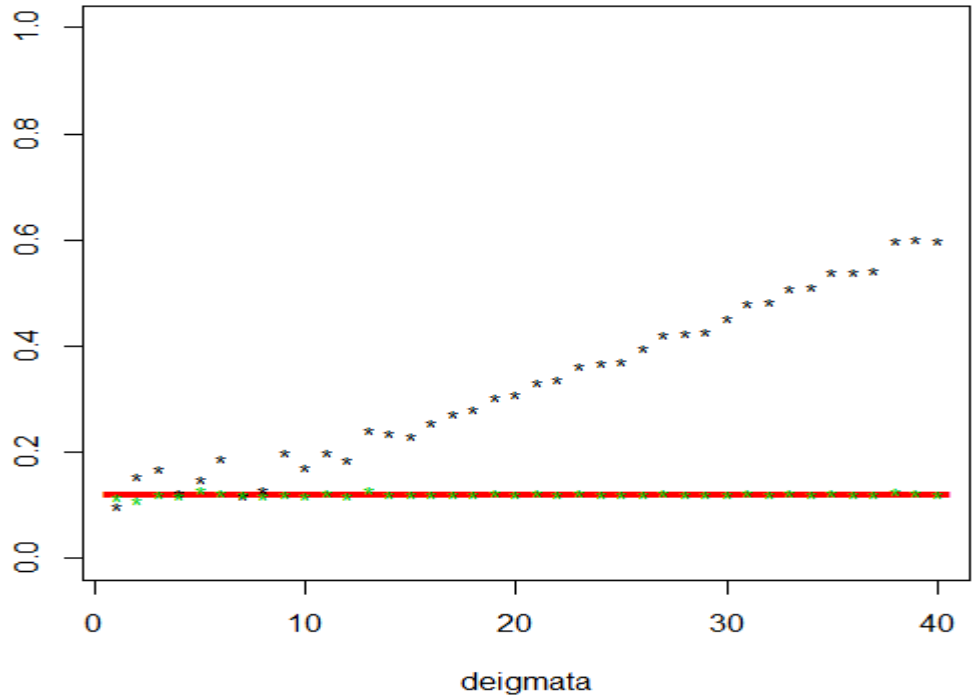
```
> PROS(30,c(100,40,50),"HGeometrical",0.02,0)

[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              3.106092                4.693908

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
        1.947391         2.185942

[[3]]
MIN MSE(CV) MAX MSE(CV)
0.001545362 0.004708497

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    1.653759e-05     5.890212e-05
```

Πίνακας 28: Αποτελέσματα της PROS για Υπεργεωμετρική με N=100, K=40, ν=50 χωρίς επανάθεση

Καλύτερα αποτελέσματα με τη μέθοδο της Πιθανοφάνειας αλλά με μικρότερη ποσοστιαία διαφορά μεταξύ των δύο μεθόδων σε σχέση με τη μέθοδο με επανάθεση.

Παρατηρούμε ότι οι τιμές του NCV και του NMLE(CV) είναι «πολύ μικρές» κάτι που δείχνει ότι πολύ γρήγορα ικανοποιούνται τα κριτήρια τερματισμού.

Συνεπώς αφήνουμε ελεύθερο το γράφημα των CVdeig, MLE(CV) και βλέπουμε ότι



Γράφημα 27 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Υπεργεωμετρική με N=100, K=40, ν=50 χωρίς επανάθεση, χωρίς κριτήριο τερματισμού

Η μέθοδος της Πιθανοφάνειας είναι αμερόληπτος εκτιμητής σε αυτή την περίπτωση (βλέπε Hanwen Zhang(2009) «A Note About Maximum Likelihood Estimator in Hyper geometric Distribution») ενώ όπως περιμέναμε η εκτίμηση μέσω του δειγματικού συντελεστή μεταβλητότητας αποκλίνει καθώς αυξάνει το δείγμα λόγω του ότι οι σχετικές συχνότητες των παρατηρήσεων μοιράζονται ομοιόμορφα.

5.8 ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΩΡΙΣ ΕΠΑΝΑΘΕΣΗ ΓΙΑ ΑΡΝΗΤΙΚΗ ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Τρέχουμε τη συνάρτηση PROS .Επιλέγουμε απόλυτο σφάλμα 0.011.

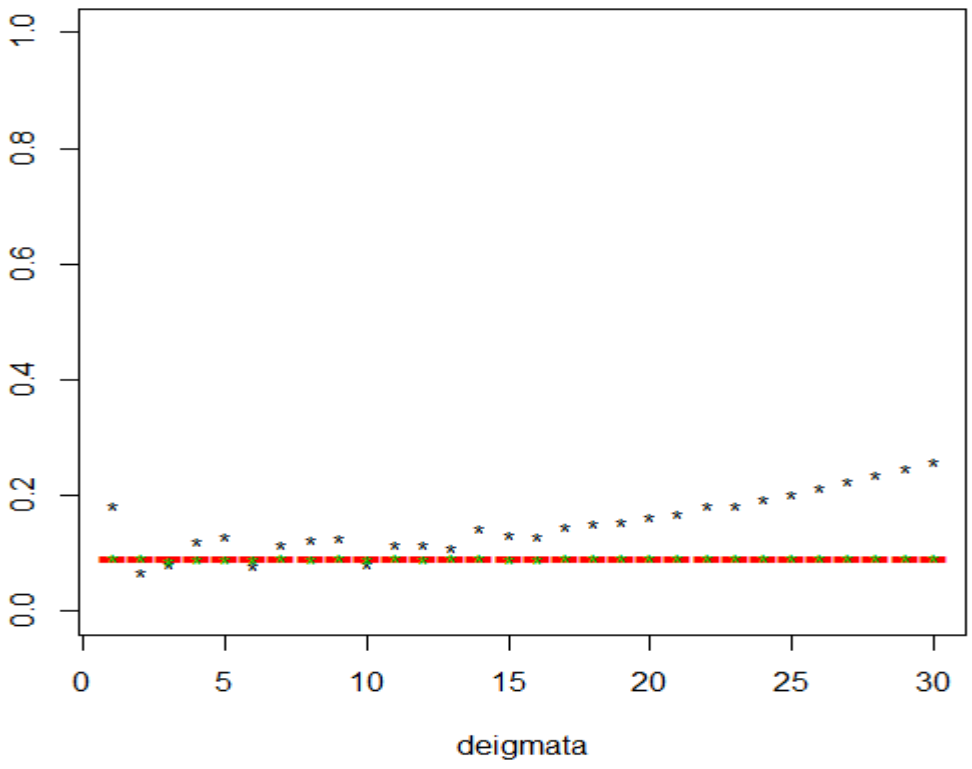
```
> PROS(100,c(20,30,60),"NHGeometrical",0.011,0)
```

```
[[1]]
MIN n me deigmatiko CV MAX n me deigmatiko CV
              4.677271                5.922729

[[2]]
MIN n me MLE(CV) MAX n me MLE(CV)
         1.98425          2.03575

[[3]]
MIN MSE(CV) MAX MSE(CV)
0.001050329 0.001966731

[[4]]
MIN MSE(MLE(CV)) MAX MSE(MLE(CV))
    3.311273e-06     7.444308e-06
```

Πίνακας 29: Αποτελέσματα της PROS για Αρνητική Υπεργεωμετρική με ν=20, K=30, N=60, χωρίς επανάθεση

Παρατηρούμε ότι οι τιμές του NCV και του NMLE(CV) είναι «πολύ μικρές» κάτι που δείχνει ότι πολύ γρήγορα ικανοποιούνται τα κριτήρια τερματισμού.

Συνεπώς αφήνουμε ελεύθερο το γράφημα των CVdeig, MLE(CV) και βλέπουμε ότι



Γράφημα 28 : CVdeig.,CV,MLE(CV) για διάφορα δείγματα χωρίς επανάθεση στην Αρνητική Υπεργεωμετρική με ν=20, K=30, N=60 χωρίς επανάθεση, χωρίς κριτήριο τερματισμού

Η μέθοδος της Πιθανοφάνειας είναι πάλι αμερόληπτος εκτιμητής και σε αυτήν την περίπτωση(βλέπε Lei Zhang και William D.Johnson (2011) «Approximate Confidence Intervals for a Parameter of the Negative Hyper geometric Distribution»**)** ενώ η εκτίμηση με τον δειγματικό συντελεστή μεταβλητότητας αποκλίνει ξανά.

ΑΝΑΚΕΦΑΛΑΙΩΣΗ

Συνοψίζοντας στο κεφάλαιο 5 θεμελιώσαμε στο προγραμματιστικό περιβάλλον της R, την θεωρία των προηγούμενων κεφαλαίων στη περίπτωση της δειγματοληψίας χωρίς επανάθεση. Τα αποτελέσματα ήταν καλύτερα με τη μέθοδο της Πιθανοφάνειας ωστόσο δεν είδαμε μεγάλες ποσοστιαίες διαφορές όπως με τη δειγματοληψία με επανάθεση. Αξιοσημείωτο είναι ότι για μικρά δείγματα πετυχαίνουμε καλύτερη εκτίμηση με τη μέθοδο της Πιθανοφάνειας ενώ καθώς το δείγμα μεγαλώνει υπάρχει απόκλιση των εκτιμήσεων CVdeig., MLE(CV) από το CV εκτός των περιπτώσεων της ομοιόμορφης, της υπεργεωμετρικής και της αρνητικής υπεργεωμετρικής. Αυτό συμβαίνει γιατί κάνοντας δειγματοληψία χωρίς επανάθεση καθώς αυξάνει το δείγμα μοιράζονται ομοιόμορφα τα «βάρη» (σχετικές συχνότητες) επιλογής των στοιχείων ομοιόμορφα κάτι που δεν ισχύει για τις περισσότερες κατανομές. Τέλος οι MLE(CV) για τις περιπτώσεις της υπεργεωμετρικής και της αρνητικής υπεργεωμετρικής κατανομής είναι αμερόληπτοι εκτιμητές.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα συμπεράσματα που προκύπτουν για κάθε τρόπο δειγματοληψίας μοιάζουν όπως είδαμε στα κεφάλαια 4 και 5, ωστόσο έχουν κάποιες σημαντικές διαφορές. Στη μέθοδο δειγματοληψίας με επανάθεση οι εκτιμητές MLE(CV) είναι καλύτεροι από τους εκτιμητές CVdeig. σε όλες τις περιπτώσεις καθώς έχουν μικρότερο μέσο τετραγωνικό σφάλμα και χρειάζονται μικρότερο δείγμα για να εκτιμήσουν τον CV με επιθυμητή ακρίβεια. Επίσης έχουν τις ιδιότητες της συνέπειας, της ασυμπτωτικής αμεροληψίας και της ασυμπτωτικής αποτελεσματικότητας που έχουν συνήθως οι Ε.Μ.Π., δηλαδή συγκλίνουν με αποτέλεσμα να υπάρχει πάντα επιθυμητό δείγμα για οποιοδήποτε απόλυτο σφάλμα που επιλέγει ο χρήστης.

Όσον αφορά την δειγματοληψία χωρίς επανάθεση οι εκτιμητές MLE(CV) είναι πάλι «καλύτεροι» από τους εκτιμητές CVdeig. για τους ίδιους λόγους όπως στη δειγματοληψία με επανάθεση (χαμηλό μέσο τετραγωνικό σφάλμα και μικρότερο πλήθος δείγματος) ,ωστόσο η βελτίωση ποσοστιαία είναι μικρότερη. Παρόλα αυτά η γενική εικόνα είναι ότι υπάρχει μεγάλη «οικονομία» στο δείγμα, όταν κάνουμε δειγματοληψία με επανάθεση είτε με τη μία προσέγγιση είτε με την άλλη, καθώς με πολύ μικρό δείγμα πετυχαίνουμε ακρίβεια η οποία όμως δεν είναι πάντοτε η επιθυμητή. Με λίγα λόγια δε παρατηρούμε γενικώς σύγκλιση, αλλά αντιθέτως στις περισσότερες περιπτώσεις καθώς αυξάνει το δείγμα οι προσεγγίσεις αποκλίνουν (εκτός των περιπτώσεων της ομοιόμορφης ,της γεωμετρικής και της αρνητικής υπεργεωμετρικής). Αυτό συμβαίνει γιατί στη δειγματοληψία χωρίς επανάθεση , καθώς αυξάνει το δείγμα , τα «βάρη» (σχετικές συχνότητες) των τιμών μοιράζονται ομοιόμορφα κάτι που δεν ισχύει σε όλες τις κατανομές. Όπως και να χει όμως ένα μικρό δείγμα μπορεί να θεωρηθεί αξιόπιστο. Μία δικλείδα ασφαλείας είναι να παίρνουμε μέγεθος δείγματος n τέτοιο ώστε να ισχύει $30<n<N-30$, όπου N το μέγεθος του πληθυσμού, Φαρμάκης Ν (2017).

Ολοκληρώνοντας αναφέρουμε ότι ενώ τα παραπάνω παραδείγματα φαίνονται ως διαδικασία προσομοίωσης, στην πραγματικότητα ο αλγόριθμος λειτουργεί ως διαδικασία εξομοίωσης. Μπορεί δηλαδή ο χρήστης να επιλέξει οποιαδήποτε διακριτή κατανομή θέλει (από αυτές που μελετήθηκαν), οποιεσδήποτε παραμέτρους (που να ικανοποιούν τους εκάστοτε μαθηματικούς περιορισμούς) και οποιοδήποτε σφάλμα. Στη περίπτωση της δειγματοληψίας με επανάθεση θα υπάρχει πάντα απάντηση για το επιθυμητό διάστημα εμπιστοσύνης του δείγματος ενώ στη χωρίς αν ο αλγόριθμος δε «τρέχει» ο χρήστης πρέπει να αυξήσει το σφάλμα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

ΦΑΡΜΑΚΗΣ Ν. (2017) «*Εισαγωγή στη Δειγματοληψία***»,** Αφοί Κυριακίδη ΕΚΔΟΣΕΙΣ Α.Ε., Θεσσαλονίκη.

ΚΟΛΥΒΑ-ΜΑΧΑΙΡΑ Φ.,ΜΠΟΡΑ-ΣΕΝΤΑ Ε.(1995) «*Στατιστική Θεωρία Εφαρμογές»*, Ζήτη ΕΚΔΟΣΕΙΣ, Θεσσαλονίκη.

ΚΟΛΥΒΑ-ΜΑΧΑΙΡΑ Φ., ΧΑΤΖΟΠΟΥΛΟΣ Σ.Α. (2015) «*Μαθηματική Στατιστική ,Έλεγχοι υποθέσεων»*, Σ.Ε.Α.Β. ΕΚΔΟΣΕΙΣ, Αθήνα.

ΚΟΥΡΟΥΚΛΗΣ Σ.-ΠΕΤΡΟΠΟΥΛΟΣ Κ. - ΠΠΕΡΙΓΚΟΥ Β.(2015) «*Θέματα Παραμετρικής Στατιστικής Συμπερασματολογίας-Εκτιμητική και Διαστήματα Εμπιστοσύνης»*, Σ.Ε.Α.Β. ΕΚΔΟΣΕΙΣ, Αθήνα.

ΚΟΥΝΙΑΣ Σ.-ΜΩΥΣΙΑΔΗΣ Χ.(1995) «*Θεωρία Πιθανοτήτων Ι Κλασική Πιθανότητα Μονοδιάστατες Κατανομές»*, Ζήτη ΕΚΔΟΣΕΙΣ, Θεσσαλονίκη.

ΚΟΥΝΙΑΣ Σ., ΚΑΛΠΑΖΙΔΟΥ Σ. (1991) «*Πιθανότητες ΙΙ,Θεωρία και Ασκήσεις*» , Ζήτη ΕΚΔΟΣΕΙΣ, Θεσσαλονίκη.

ΡΟΥΣΣΑΣ Γ.(1997) «*Στατιστική Συμπερασματολογία, Τόμος ΙΙ Έλεγχος υποθέσεων* », Ζήτη ΕΚΔΟΣΕΙΣ, Θεσσαλονίκη.

ΖΑΧΑΡΟΠΟΥΛΟΥ Χ.(2015) «*Στατιστική ,Μέθοδοι – Εφαρμογές, Τόμος Α, 6η Έκδοση»* Σοφία ΕΚΔΟΣΕΙΣ Α.Ε., Θεσσαλονίκη.

ΚΥΒΕΝΤΙΔΗΣ Θ.(2001) «*Διαφορικός Λογισμός Συναρτήσεων Μιας Πραγματικής Μεταβλητής, Τεύχος Πρώτο»*, Ζήτη ΕΚΔΟΣΕΙΣ, Θεσσαλονίκη.

ΑΝΔΡΕΑΔΑΚΗΣ Σ.-ΚΑΤΣΑΡΓΥΡΗΣ Β.- ΠΑΠΑΣΤΑΥΡΙΔΗΣ Σ.-.ΠΟΛΥΖΟΣ Γ. - ΣΒΕΡΚΟΣ Α.- ΑΔΑΜΟΠΟΥΛΟΣ Λ.- ΔΑΜΙΑΝΟΥ Χ.(1998) «*Άλγεβρα και στοιχεία πιθανοτήτων Α' Γενικού Λυκείου»*, ΙΝΣΤΙΤΟΥΤΟ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΕΚΔΟΣΕΩΝ «ΔΙΟΦΑΝΤΟΣ».

ΠΑΠΑΝΤΩΝΗΣ Ν.Γ.(2016) «*Η αμεροληψία του Συντελεστή Μεταβλητότητας και με ποιους τρόπους επιτυγχάνεται»*

ΞΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

R.MAHMOUDVAND, H.HASSANI (2007) "Is The Sample Coefficient Of Variation A Good Estimator For The Population Coefficient Of Variation?", *World Applied Sciences Journal 2*, (5), pp: 519-522, IDOSI Publications.

R.BREUNIG (2001), "An almost unbiased estimator of the coefficient of variation", *Economics letters 70*, pp :15-19

Nairy, K.S. & Rao, K.N. (2003). "Tests of coefficients of variation of normal populations", *Communications in Statistics Simulation and Computation,* 32, pp:21-32.

ΔΙΑΔΥΚΤΙΟ

https://en.wikipedia.org/wiki/Negative_hypergeometric_distribution

http://www.math.ntua.gr/~fouskakis/exercises-stats.pdf

http://users.uoi.gr/alapatin/files/Lecture%201_presentation.pdf