Κεφάλαιο 7:
Η επιλογή των πιθανοτικών κατανομών εισόδου

Τεχνικές Εκτίμησης Υπολογιστικών συστημάτων

Γιάννης Γαροφαλάκης

Καθηγητής

Εισαγωγή

- Υλοποίηση μιας προσομοίωσης με τη χρήση τυχαίων εισόδων
  - Καθορισμός πιθανοτικών κατανομών που τις περιγράφουν
  - Η προσομοίωση εξελίσσεται με τη δημιουργία και χρήση *τυχαίων τιμών* από τις κατανομές αυτές

Εισαγωγή (2)

- Πηγές Τυχαιότητας Πραγματικών Συστημάτων

Τύπος Συστήματος	*Πηγές Τυχαιότητας*
Υπολογιστές	**Χρόνοι μεταξύ αφίξεων εργασιών,** **Τύποι εργασιών,** **Απαιτήσεις επεξεργασίας εργασιών.** **Επιλογή σελίδων κ.α. στο Web**
Δίκτυα - Τηλεπικοινωνίες	**Χρόνοι μεταξύ αφίξεων μηνυμάτων,** **Τύποι μηνυμάτων,** **Μήκη μηνυμάτων.**
Συστήματα Παραγωγής	**Χρόνοι επεξεργασίας,** **Χρόνοι λειτουργίας μηχανών μέχρι να χαλάσουν,** **Χρόνοι επισκευής μηχανών.**

Εισαγωγή – Συλλογή δεδομένων

- Αν είναι δυνατόν να συλλέξουμε δεδομένα από μια τυχαία μεταβλητή εισόδου που μας ενδιαφέρει, μπορούμε να τα χρησιμοποιήσουμε με τους εξής τρόπους, προκειμένου να προσδιορίσουμε την κατανομή εισόδου
  - Τα ίδια τα δεδομένα χρησιμοποιούνται απευθείας στην προσομοίωση. Τότε θα έχουμε *ιχνο-οδηγούμενη προσομοίωση.*
  - Τα δεδομένα χρησιμοποιούνται για να ορισθεί μία *εμπειρική* συνάρτηση κατανομής για τη μεταβλητή εισόδου.
  - Προσαρμόζουμε μια γνωστή *θεωρητική* μορφή κατανομής (π.χ. εκθετική ή Poisson) στα δεδομένα που έχουμε.

Εισαγωγή – Συλλογή δεδομένων (2)

- Σύγκριση
  - 1ος τρόπος
    - Η προσομοίωση θα αναπαράγει απλώς ότι έγινε στην πραγματικότητα.
    - Σπανίως έχουμε αρκετά δεδομένα για να εκτελέσουμε όλες τις απαραίτητες προσομοιώσεις.
  - 2ος τρόπος
    - Αντιμετωπίζει τα μειονεκτήματα αυτά, αφού τουλάχιστον για συνεχή δεδομένα, μπορεί να παραχθεί οποιαδήποτε τιμή μεταξύ της ελάχιστης και μέγιστης που έχουμε στη διάθεσή μας.
  - Είναι προτιμητέος ο δεύτερος από τον πρώτο τρόπο, ο πρώτος συνιστάται για τον *έλεγχο εγκυρότητας ενός μοντέλου,* συγκεκριμένα όταν θέλουμε να συγκρίνουμε τις εξόδους του μοντέλου με τις εξόδους του πραγματικού συστήματος.

Εισαγωγή – Συλλογή δεδομένων (3)

- Αν μπορούμε να βρούμε μια θεωρητική κατανομή που προσαρμόζεται ικανοποιητικά στα δεδομένα που έχουμε (τρίτος τρόπος), την προτιμούμε από την εμπειρική κατανομή (δεύτερος τρόπος), για τους παρακάτω λόγους:

Εισαγωγή – Χαρακτηριστικά κατανομών

- Μία εμπειρική κατανομή μπορεί να έχει σημεία ασυνέχειας, ή άλλες “ανωμαλίες”, ιδιαίτερα αν διαθέτουμε λίγα μόνο δεδομένα για τον προσδιορισμό της. Αντίθετα, μια θεωρητική κατανομή εξομαλύνει τα δεδομένα και μπορεί να δώσει πληροφορία για τη συνολική συμπεριφορά της μεταβλητής εισόδου.
- Συνήθως δεν μπορούμε να παράγουμε τυχαίες τιμές από μια εμπειρική κατανομή έξω από τα όρια των δεδομένων που διαθέτουμε. Το γεγονός αυτό δημιουργεί μερικές φορές σημαντικό πρόβλημα, αφού η απόδοση ενός συστήματος μπορεί να εξαρτάται από την πιθανότητα εμφάνισης ενός ακραίου γεγονότος, όπως για παράδειγμα ενός πολύ μεγάλου χρόνου εξυπηρέτησης. Αντίθετα, μια θεωρητική κατανομή μπορεί να μας δώσει και τιμές έξω από το μέγιστο και ελάχιστο των δεδομένων που διαθέτουμε.

Χρήσιμες πιθανοτικές κατανομές

- *Εκθετική (exponential):* Χρησιμοποιείται για τη μοντελοποίηση εντελώς “τυχαίων” γεγονότων, όπως οι χρόνοι μεταξύ αστοχιών, ή οι χρόνοι μεταξύ αφίξεων πελατών. Συχνά χρησιμοποιείται ως η μόνη λύση όταν δεν υπάρχουν ενδείξεις για την κατανομή που πρέπει να χρησιμοποιηθεί.
- *Βήτα (beta):* Χρησιμοποιείται για τη μοντελοποίηση τυχαίων αναλογιών, όπως το κλάσμα των πακέτων που απαιτούν επανεκπομπή σε μία γραμμή μετάδοσης δεδομένων.
- *Zipf:* Για την επιλογή από σύνολο σελίδων στο Web, από σύνολο multimedia αντικειμένων (π.χ. ταινίες σε video on demand σύστημα) κ.α. Τα στοιχεία του συνόλου είναι ταξινομημένα με βάση τη δημοτικότητα.
- *Γάμμα (gamma):* Χρησιμοποιείται για τη μοντελοποίηση χρόνων εξυπηρέτησης. Σε ένα σύστημα αναμονής περιγράφει το χρόνο εξυπηρέτησης, όταν έχουμε m εκθετικούς εξυπηρετητές εν σειρά.
- *Δυωνυμική (binomial):* Χρησιμοποιείται στη μοντελοποίηση του αριθμού επιτυχιών σε μια ακολουθία *n* ανεξαρτήτων δοκιμών Bernoulli. Για παράδειγμα, ο αριθμός των bits σε ένα πακέτο που δεν έχουν επηρεασθεί από το θόρυβο.

Χρήσιμες πιθανοτικές κατανομές (2)

- *Γεωμετρική (geometric):* Χρησιμοποιείται στη μοντελοποίηση του αριθμού των εμφανίσεων γεγονότων μεταξύ άλλων σημαντικών γεγονότων. Π.χ. ο αριθμός των σωστών bits μεταξύ δύο λανθασμένων bits σε ένα πακέτο δεδομένων, ή ο αριθμός των αιτήσεων σε μια τοπική Βάση Δεδομένων (ΒΔ) μεταξύ δύο αιτήσεων σε μια απομακρυσμένη ΒΔ.
- *Αρνητική Δυωνυμική (negative binomial):* Χρησιμοποιείται για τον αριθμό των αποτυχιών πριν τη *m*-στή επιτυχία, όπως ο αριθμός των επανεκπομπών ενός αρχείου που αποτελείται από *m* πακέτα, ή ο αριθμός των αιτήσεων σε μια τοπική ΒΔ πριν τη *m*-στή αίτηση σε μια απομακρυσμένη ΒΔ.
- *Poisson:* Χρησιμοποιείται στη μοντελοποίηση του αριθμού εμφανίσεων γεγονότων κατά τη διάρκεια μιας συγκεκριμένης περιόδου, όπως ο αριθμός των αστοχιών εξαρτημάτων στη μονάδα του χρόνου, ή ο αριθμός των αιτήσεων σε μια ΒΔ σε χρονικό διάστημα *t.*
- *Κανονική (normal):* Χρησιμοποιείται για τη μοντελοποίηση του αθροιστικού φαινομένου διαφόρων ανεξαρτήτων πηγών, όπως τα λάθη μετρήσεων.
- *Weibull:* Χρησιμοποιείται για μετρήσεις αξιοπιστίας, όπως ο χρόνος ζωής εξαρτημάτων.

Εμπειρικές κατανομές

- Όταν δεν είναι δυνατόν να βρούμε μια θεωρητική κατανομή που να ταιριάζει στα δεδομένα που διαθέτουμε, προσπαθούμε να καθορίσουμε απευθείας μια *εμπειρική κατανομή,* από την οποία θα δημιουργούνται τυχαίες τιμές κατά τη διάρκεια της προσομοίωσης.
- Για **συνεχείς** τυχαίες μεταβλητές, ο τύπος της εμπειρικής κατανομής που μπορούμε να ορίσουμε, εξαρτάται από το αν διαθέτουμε τις ίδιες τις τιμές των επιμέρους *αυθεντικών* παρατηρήσεων $X_1, X_2, \ldots, X_n$, ή μόνο τον *αριθμό* των X_i που βρίσκονται μέσα σε κάθε ένα από ορισμένα καθορισμένα διαστήματα.
- Στη δεύτερη περίπτωση λέμε ότι έχουμε *ομαδοποιημένα* δεδομένα, ή δεδομένα με τη μορφή *ιστογράμματος.*

Εμπειρικές κατανομές – Αυθεντικά δεδομένα

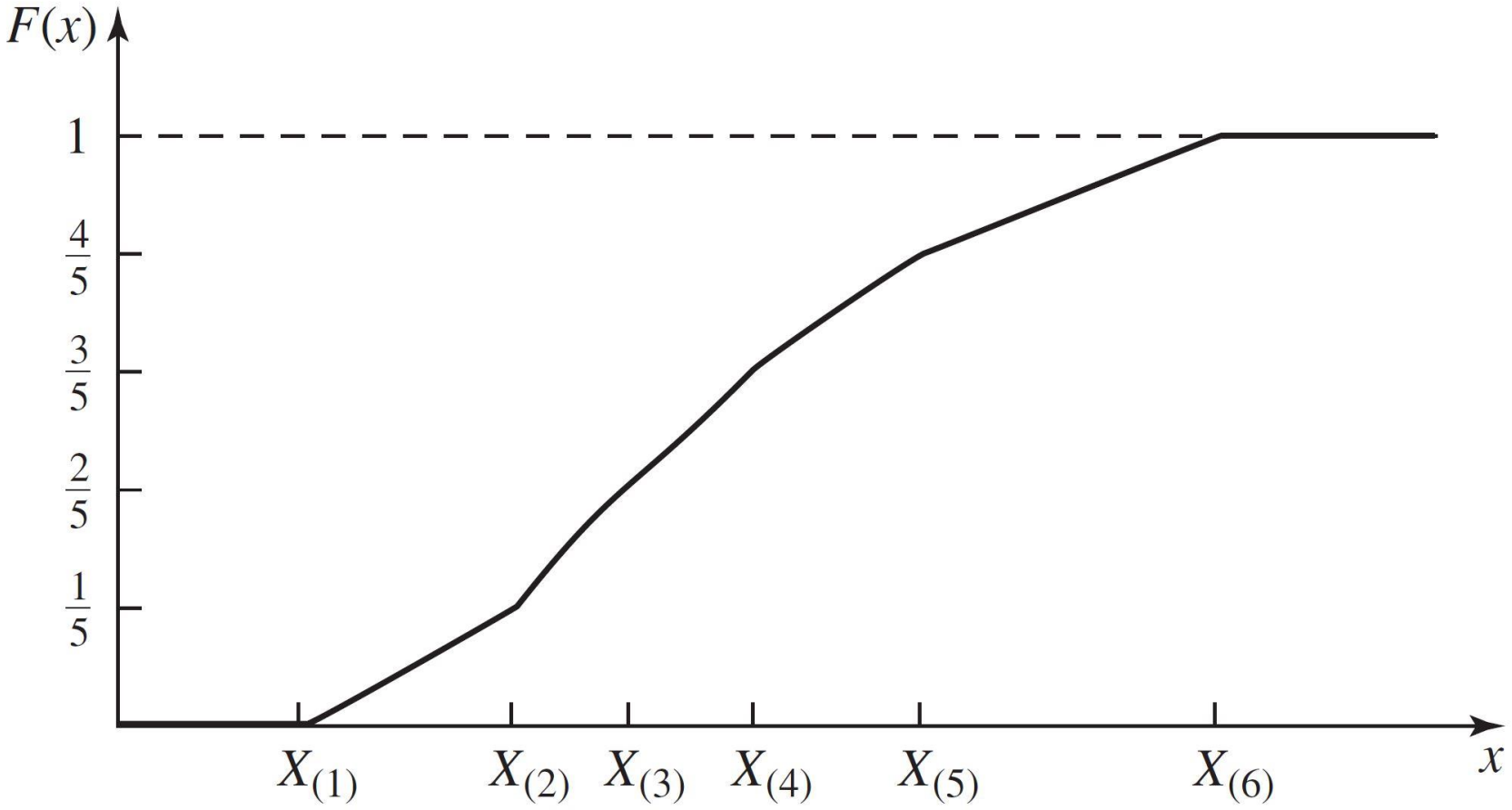
- Αν έχουμε στη διάθεσή μας τα αυθεντικά δεδομένα, μπορούμε να ορίσουμε μία συνεχή, τμηματικά γραμμική, συνάρτηση κατανομής *F,* έχοντας ταξινομήσει τα X κατά αύξουσα σειρά.
- Έστω $X_{(i)}$ το *i*-στό μικρότερο από τα X_j έτσι ώστε $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

Εμπειρικές κατανομές – Αυθεντικά δεδομένα (2)

- η *F* δίνεται από τη σχέση:

$$F(x)=\begin{cases} 0 & \alpha\nu \quad x < X_{(1)} \\ \dfrac{i-1}{n-1}+\dfrac{x-X_{(i)}}{(n-1)(X_{(i+1)}-X_{(i)})} & \alpha\nu \quad X_{(i)} \le x \le X_{(i+1)} \quad \gamma\iota\alpha \quad i=1,2,\ldots,n-1 \\ 1 & \alpha\nu \quad X_{(n)} \le x \end{cases}$$

Εμπειρικές κατανομές – Αυθεντικά δεδομένα (3)



Συνεχής, τμηματικά γραμμική, εμπειρική συνάρτηση κατανομής από τα αυθεντικά δεδομένα

Εμπειρικές κατανομές – Αυθεντικά δεδομένα (4)

- Το Σχήμα 7.1 δείχνει μία περίπτωση για $n = 6$
- Το $F(x)$ αυξάνεται γρηγορότερα στις περιοχές του *x* στις οποίες τα X_i είναι κατανεμημένα πιο πυκνά, όπως επιθυμούμε.
- Για κάθε *i*, $F(X_{(i)}) = (i-1)/(n-1)$, το οποίο είναι κατά προσέγγιση (για μεγάλο *n*) το ποσοστό των X_i που είναι *μικρότερα από* $X_{(i)}$ γεγονός επίσης επιθυμητό για μία *συνεχή* συνάρτηση κατανομής.
- Ένα μειονέκτημα της εμπειρικής κατανομής που φτιάξαμε με τη μέθοδο αυτή, είναι ότι οι τυχαίες τιμές που δημιουργούνται από αυτήν κατά τη διάρκεια της προσομοίωσης, δεν μπορούν να είναι μικρότερες από $X_{(1)}$ ή μεγαλύτερες από $X_{(n)}$.
- Η μέση τιμή της $F(x)$ δεν είναι ίση με τη δειγματοληπτική μέση τιμή $\bar{X}(n)$ των X_i .

Εμπειρικές κατανομές – Ομαδοποιημένα δεδομένα

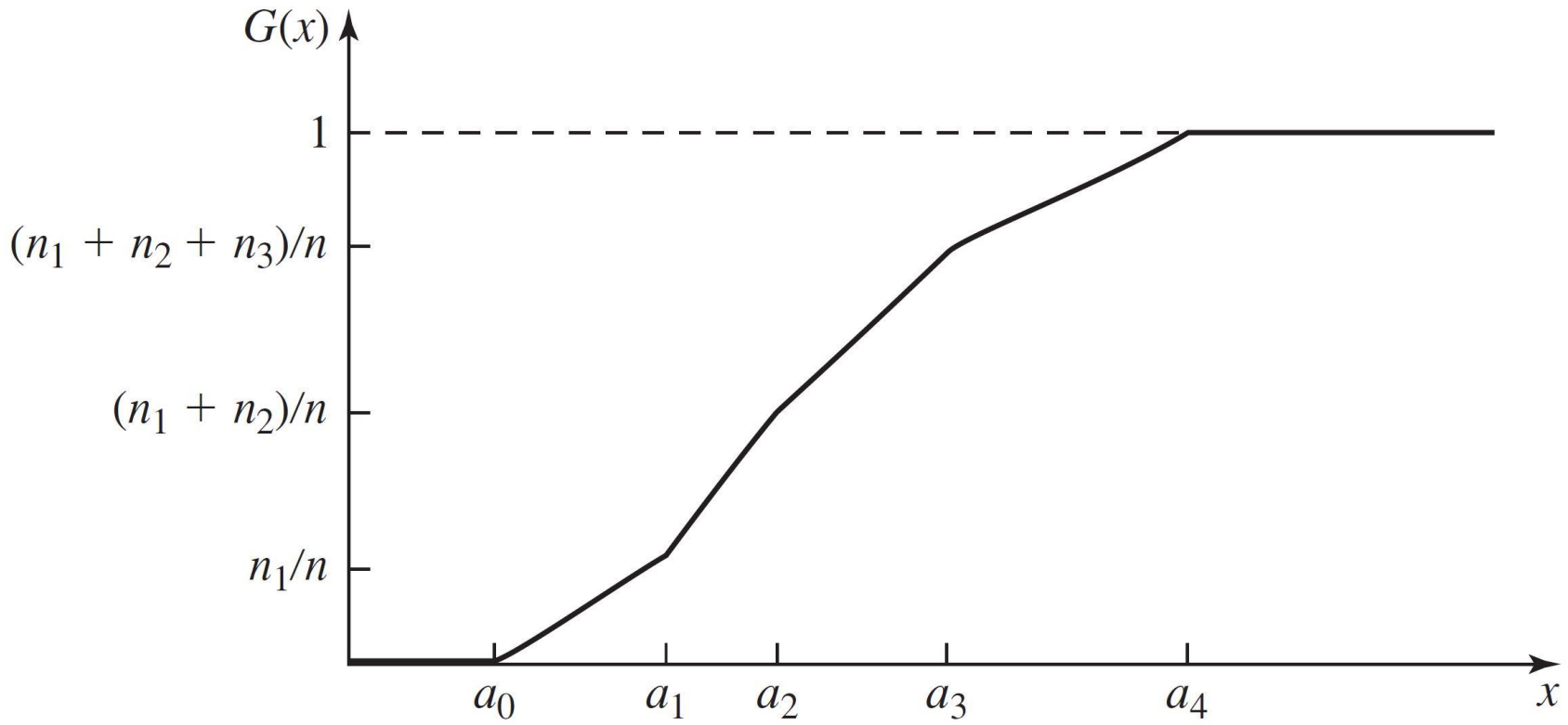
- Διαφορετική προσέγγιση
- Δεν γνωρίζουμε τις τιμές των επιμέρους X_i
- Έστω ότι τα *n* X_i είναι ομαδοποιημένα σε *k* συνεχόμενα διαστήματα $[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k)$, έτσι ώστε το *j*-στό διάστημα να περιέχει n_j παρατηρήσεις, όπου $n_1 + n_2 + \cdots + n_k = n$.
- Μία λογική, τμηματικά γραμμική εμπειρική συνάρτηση κατανομής *G*
  - $G(a_0) = 0$
  - $G(a_j) = (n_1 + n_2 + \cdots + n_j) / n$ για $j = 1, 2, \ldots, k$

Εμπειρικές κατανομές – Ομαδοποιημένα δεδομένα (2)

- Με γραμμική παρεμβολή μεταξύ των a_j

$$G(x) = \begin{cases} 0 & \alpha\nu \quad x < a_0 \\ G(a_{j-1}) + \dfrac{x - a_{j-1}}{a_j - a_{j-1}}\left[G(a_j) - G(a_{j-1})\right] & \alpha\nu \quad a_{j-1} \le x < a_j \quad \gamma\iota\alpha \;\; j = 1,2,\ldots,k \\ 1 & \alpha\nu \quad a_k \le x \end{cases}$$

Εμπειρικές κατανομές – Ομαδοποιημένα δεδομένα (3)



Συνεχής, τμηματικά γραμμική, εμπειρική συνάρτηση κατανομής από ομαδοποιημένα δεδομένα

Εμπειρικές κατανομές – Ομαδοποιημένα δεδομένα (4)

- Στο Σχήμα 7.2 φαίνεται μία τέτοια κατανομή $G(x)$ για $k = 4$.
- Το $G(a_j)$ είναι το ποσοστό των X_i που είναι μικρότερα από a_j και η $G(x)$ αυξάνει γρηγορότερα στα διαστήματα του *x* όπου οι παρατηρήσεις είναι πιο πυκνές.
- Οι τυχαίες τιμές που θα δημιουργηθούν από την κατανομή αυτή, θα βρίσκονται μεταξύ των a_0 και a_k .

Εμπειρικές κατανομές – Ομαδοποιημένα δεδομένα (5)

- Στην πράξη, πολλές συνεχείς κατανομές εκτείνονται μειούμενες προς τα δεξιά, με μία μορφή συνάρτησης πυκνότητας πιθανότητας $f(x)$ παρόμοια με αυτήν της κατανομής gamma.
- Αν το μέγεθος του δείγματος *n* δεν είναι πολύ μεγάλο, μάλλον θα έχουμε λίγες (αν έχουμε) παρατηρήσεις από τη δεξιά ουρά της πραγματικής κατανομής, αφού η πιθανότητα να βρεθούμε στην ουρά είναι συνήθως μικρή.

Εμπειρικές κατανομές – Ομαδοποιημένα δεδομένα (6)

- Οι παραπάνω εμπειρικές κατανομές δεν επιτρέπουν τη δημιουργία τυχαίων τιμών μεγαλύτερων από τη μέγιστη παρατήρηση.
- Οι πολύ μεγάλες τιμές μπορεί να έχουν σημαντική επίπτωση στην προσομοίωση, όπως για παράδειγμα, ένας μεγάλος χρόνος εξυπηρέτησης ο οποίος σε ένα σύστημα αναμονής πιθανόν να προκαλέσει τη δημιουργία μεγάλων ουρών και κατά συνέπεια καθυστερήσεων.
  - Έχει προταθεί να προστίθεται μία εκθετική κατανομή στο δεξιό τμήμα της εμπειρικής κατανομής, ώστε να επιτρέπεται η δημιουργία τυχαίων τιμών μεγαλύτερων από $X_{(n)}$

Εμπειρικές κατανομές – Διακριτά δεδομένα

- Είναι αρκετά εύκολο να ορισθεί μία εμπειρική κατανομή, αν είναι διαθέσιμες οι αυθεντικές τιμές $X_1, X_2, \ldots, X_n$.
- Για κάθε πιθανή τιμή *x*, μπορεί να ορισθεί μία εμπειρική συνάρτηση μάζας πιθανότητας $p(x)$ η οποία θα είναι το ποσοστό των X_i που θα είναι ίσα με *x.*
- Για ομαδοποιημένα διακριτά δεδομένα, θα μπορούσαμε να ορίσουμε μία συνάρτηση μάζας, έτσι *ώστε* το *άθροισμα* των $p(x)$ πάνω σε όλες τις πιθανές τιμές του *x* σε ένα διάστημα, να είναι ίσο με το ποσοστό των X_i στο διάστημα αυτό.
- Ο τρόπος που δημιουργούνται τα επιμέρους $p(x)$ για τις πιθανές τιμές του *x* μέσα σε ένα διάστημα, είναι ουσιαστικά αυθαίρετος.

Προσαρμογή θεωρητικής κατανομής

- Όταν θέλουμε να βρούμε μια γνωστή κατανομή και να την προσαρμόσουμε στα δεδομένα που διαθέτουμε, ώστε να τη χρησιμοποιήσουμε στη συνέχεια για τη δημιουργία τυχαίων τιμών, ακολουθούμε γενικά τα παρακάτω βήματα:
  - Δημιουργούμε ένα **ιστόγραμμα** από τα δεδομένα που διαθέτουμε.
  - Επιλέγουμε μια **γνωστή κατανομή**, με μορφή παρόμοια με αυτήν του ιστογράμματος που έχουμε κατασκευάσει.
  - Υπολογίζουμε την πρώτη και δεύτερη δειγματοληπτική **ροπή** των δεδομένων που έχουμε.
  - Χρησιμοποιούμε τις ροπές για να υπολογίσουμε τις **παραμέτρους** της θεωρητικής κατανομής (δηλ. αντιστοιχούμε τις δειγματοληπτικές ροπές των δεδομένων, με τις ροπές της θεωρητικής κατανομής).
  - Κάνουμε ένα **τεστ** για να δούμε πόσο καλά αντιπροσωπεύει τα δεδομένα μας, η θεωρητική κατανομή που επιλέξαμε (π.χ. το χ^2 τεστ για διακριτές κατανομές και το τεστ Kolmogorov - Smirnov για συνεχείς κατανομές).

Προσαρμογή θεωρητικής κατανομής (2)

- Αν δεν μπορέσουμε να βρούμε θεωρητική κατανομή με τον τρόπο αυτό, τότε ίσως να είναι δυνατόν να μετασχηματίσουμε με κάποιο τρόπο τα αρχικά δεδομένα.
- Δημιουργούμε ένα νέο σύνολο δεδομένων Y_i από τα αρχικά δεδομένα X_i , με $Y_i = g(X_i)$, όπου $g(x)$ είναι μία συνάρτηση μετασχηματισμού.
- Εφαρμόζουμε και πάλι τα παραπάνω 5 βήματα, ελπίζοντας να προσδιορίσουμε μία θεωρητική κατανομή που θα προσαρμοσθεί ικανοποιητικά στα μετασχηματισμένα δεδομένα Y_i .
- Στην επιλογή της $g(x)$ συνιστάται να χρησιμοποιούμε απλές συναρτήσεις, όπως γραμμικούς μετασχηματισμούς, λογαρίθμους, ή δυνάμεις. Είναι πιθανό πάντως να χρειασθούμε αρκετές δοκιμές μέχρι να βρούμε μια θεωρητική κατανομή που να προσαρμόζεται ικανοποιητικά στα αρχικά δεδομένα.