Πανεπιστήμιο Αιγαίου

Χωρική Ανάλυση

Ενότητα 5γ: Τυχαίες μεταβλητές & στατιστικές κατανομές

Κυριακίδης Φαίδων

Τμήμα Γεωγραφίας

Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.

Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Τυχαίες Μεταβλητές και Στατιστικές Κατανομές

Φαίδων Κυριακίδης

Καθηγητής

phkyriakidis@geo.aegean.gr

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
Λόφος Πανεπιστημίου, 81100 Μυτιλήνη

Χωρική Ανάλυση

ΤΜΗΜΑ ΓΕΩΓΡΑΦΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΓΕΩΓΡΑΦΙΑ ΚΑΙ ΕΦΑΡΜΟΣΜΕΝΗ ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ

Τυχαίες Μεταβλητές και Κατανομές

Περιγραφική στατιστική ανάλυση και συμπερασματολογία

Ενώ η περιγραφική στατιστική ασχολείται με την αποτελεσματική περιγραφή (απεικόνιση, διερεύνηση και περίληψη) δεδομένων/μετρήσεων, ο απώτερος στόχος της στατιστικής ανάλυσης είναι η εξαγωγή συμπερασμάτων σχετικά με τα χαρακτηριστικά και τις ιδιότητες ενός πληθυσμού με βάση τις μετρήσεις που περιλαμβάνονται σε ένα δείγμα

Πληθυσμός και δείγμα

Βασικό εργαλείο στην εξαγωγή συμπερασμάτων για τα χαρακτηριστικά μιας μεταβλητής που περιγράφει έναν πληθυσμό είναι η μοντελοποίηση της κατανομής της, δηλαδή της συχνότητας εμφάνισης των διαφόρων τιμών της μεταβλητής. Στο σημείο αυτό υπεισέρχεται η έννοια της θεωρητικής στατιστικής κατανομής που χρησιμεύει ως αφαιρετικό μοντέλο για την μεταβλητή και τον πληθυσμό, σε αντίθεση με τη δειγματική ή εμπειρική κατανομή που χρησιμοποιείται ως εφαλτήριο για την επιλογή μιας θεωρητικής κατανομής και την εκτίμηση των παραμέτρων της

Στόχοι του μαθήματος αυτού

- εξοικίωση με τις βασικές έννοιες που σχετίζονται με τις στατιστικές κατανομές ως μοντέλα κατανομής συχνοτήτων τυχαίων μεταβλητών
- σύντομη επισκόπηση διαφόρων τύπων στατιστικών κατανομών που υπεισέρχονται στη στατιστική ανάλυση χωρικών δεδομένων

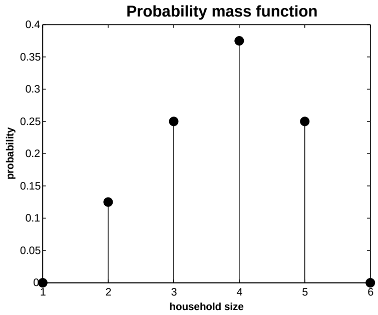
Ορισμοί

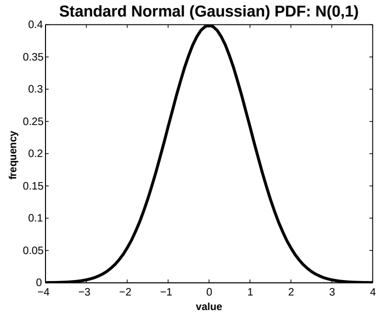
Τυχαία μεταβλητή (ΤΜ)

Μια μεταβλητή X που μπορεί να πάρει μια σειρά τιμών ή πραγματοποιήσεων, $\{x_i, i = 1, \ldots, N\}$ για ένα πληθυσμό, το N μπορεί να είναι άπειρο $N \to \infty$, για ένα δείγμα, το N είναι πεπερασμένο

Στατιστική κατανομή

Ένας πίνακας ή μια μαθηματική συνάρτηση, που αντιστοιχεί τιμές μιας μεταβλητής με πιθανότητες εμφάνισής τους





Διακριτή ή συνεχής ΤΜ

- διακριτή: ΤΜ που μπορεί να πάρει ένα πεπερασμένο σύνολο διακριτών τιμών, π.χ., απαριθμητικές μεταβλητές: πληθυσμός, αριθμός ατυχημάτων...
- συνεχής: ΤΜ που μπορεί να πάρει άπειρες τιμές μέσα σε κάποιο διάστημα, π.χ., θερμοκρασία, ταχύτητα, απόσταση...

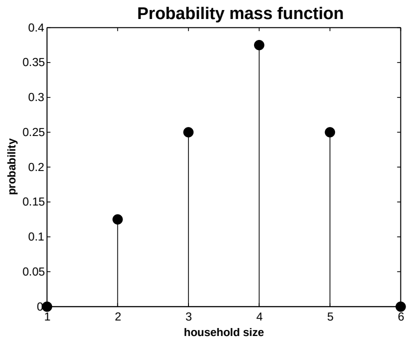
Συνάρτηση Μάζας Πιθανότητας

Ορισμός

Συνάρτηση που αντιστοιχεί σε κάθε τιμή x_i μιας **διακριτής** ΤΜ X την πιθανότητα $f(x_i)$ εμφάνισης της τιμής αυτής: $f(x_i) = Prob\{X = x_i\}$, με $f(x_i) \geq 0$ και $\sum f(x_i) = 1$.

Π.χ., η πιθανότητα εμφάνισης ενός αριθμού μελών οικογένειας στα νοικοκυριά μιας χώρας:

x_i	1	2	3	4	5	6
$f(x_i)$	0	0.125	0.250	0.375	0.250	0

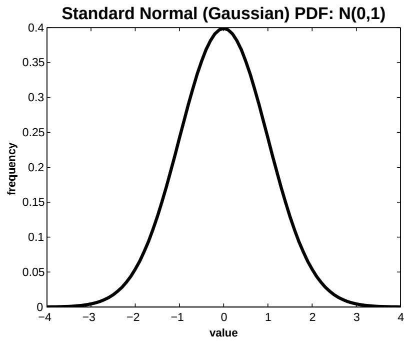


Συνάρτηση Πυκνότητας Πιθανότητας

Ορισμός ΣΠΠ

Συνάρτηση $f_X(x)$ η οποία δίνει την πιθανότητα εμφάνισης τιμών μιας **συνεχούς** τυχαίας μεταβλητής X μέσα σε ένα απειροελάχιστο διάστημα $[x - \epsilon, x + \epsilon]$:

$$f_X(x) = Prob\{X \in [x - \epsilon, x + \epsilon]\}$$



Σημείωση

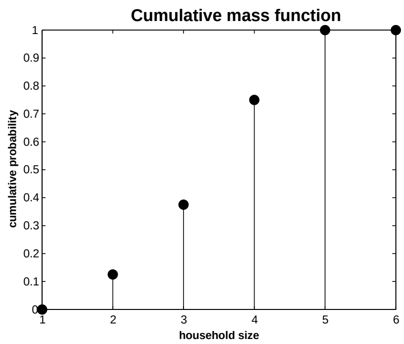
Το εμβαδόν κάτω από την καμπύλη της ΣΠΠ είναι πάντα 1: $\int_{-\infty}^{+\infty} f_X(x)dx = 1$

Συνάρτηση Αθροιστικής Μάζας

Ορισμός

Συνάρτηση που αντιστοιχεί σε κάθε τιμή x_i μιας **διακριτής** ΤΜ X την πιθανότητα $F(x_i)$ εμφάνισης τιμών μικρότερων ή ίσων της τιμής x_i: $F(x_i) = Prob\{X \leq x_i\} = \sum_{j=1}^{i} f(x_j)$

x_i	1	2	3	4	5	6
$F(x_i)$	0.000	0.125	0.375	0.750	1.000	1.000



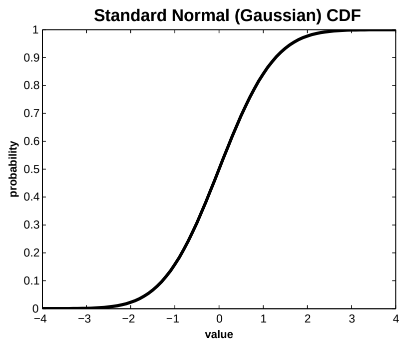
Μια συνάρτηση αθροιστικής μάζας δεν μπορεί να είναι φθίνουσα

Συνάρτηση Κατανομής

Ορισμός ΣΚ

Συνάρτηση $F_X(y)$ η οποία δίνει την πιθανότητα εμφάνισης τιμών x μιας **συνεχούς** τυχαίας μεταβλητής X, οι οποίες είναι μικρότερες ή ίσες της τιμής y:

$$F_X(y) = Prob\{X \leq y\} = \int_{-\infty}^{y} f_X(x)\, dx$$



Σημειώσεις

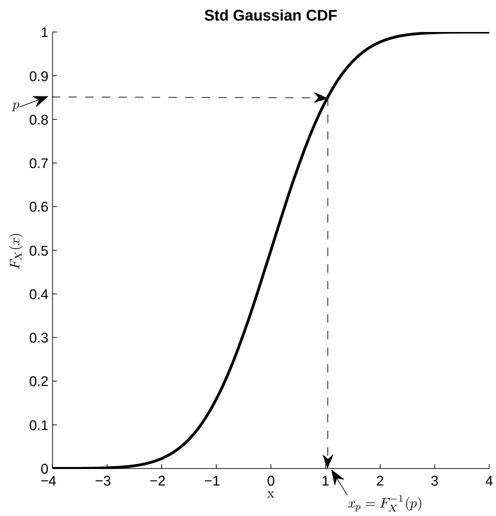
- η (ποτέ φθίνουσα) ΣΚ είναι το ολοκλήρωμα της ΣΠΠ $f_X(x)$ για όλες τις τιμές $x \leq y$, ή αλλιώς η ΣΠΠ (όχι πάντα ορίσιμη) είναι η παράγωγος της ΣΠ
- πιθανότητα εμφάνισης τιμών στο διάστημα $(a, b]$: $f(X \in (a, b])$ $= Prob\{a < X \leq b\} = Prob\{X \leq b\} - Prob\{X \leq a\} = F_X(b) - F_X(a)$

Συνάρτηση Ποσοστημορίων

Αντίστροφη συνάρτηση κατανομής

Συνάρτηση που αντιστοιχεί σε κάθε πιθανότητα p την τιμή της μεταβλητής x_p, για την οποία υπάρχει πιθανότητα p να μην εμφανιστούν τιμές μεγαλύτερες της – η τιμή x_p λέγεται το p-ποσοστημόριο:

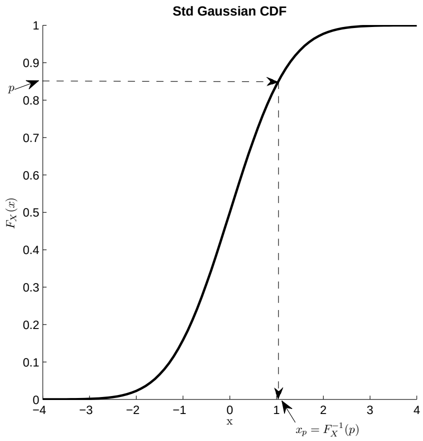
$$x_p = F_X^{-1}(p), \quad \text{όπου } p = F_X(x_p)$$



Τυχαία Δειγματοληψία

Διαδικασία δειγματοληψίας από μια συνεχή ΣΚ

1. προσομοίωση ενός τυχαίου αριθμού p από την ομοιόμορφη κατανομή στο διάστημα $[0, 1]$
2. υπολογισμός του αντίστοιχου ποσοστημόριου: $x_p = F_X^{-1}(p)$
3. επανάληψη των παραπάνω βημάτων για την παραγωγή S προσομοιωμένων τιμών $\{x_{p_i}, i = 1, \ldots, S\}$



- οι προσομοιωμένες τιμές ακολουθούν την κατανομή $F_X(x)$ τόσο πιστότερα όσο μεγαλώνει ο αριθμός S
- οι προσομοιωμένες τιμές χρησιμοποιούνται για τον υπολογισμό της αβεβαιότητας στα εξερχόμενα (outputs) μοντέλων και την ανάλυση των επιπτώσεων της

Αναμενόμενη Τιμή $E\{X\}$

Ορισμός

Μέση τιμή $E\{X\} = \mu$ μιας τυχαίας μεταβλητής X, ή αλλιώς σταθμισμένο – με βάση τη ΣΠΠ $f_X(x)$ – άθροισμα ή ολοκλήρωμα όλων των πιθανών τιμών x μιας ΤΜ X

Για διακριτές τυχαίες μεταβλητές

Χρησιμοποιώντας K πιθανές τιμές $\{x_1, \ldots, x_K\}$:

$$E\{X\} = \mu = \sum_{k=1}^{K} x_k f_X(x_k)$$

Η αναμενόμενη τιμή μπορεί να μην ανήκει στο σύνολο των πραγματοποιήσιμων τιμών

Για συνεχείς τυχαίες μεταβλητές

$$E\{X\} = \mu = \int_{-\infty}^{+\infty} x\, f_X(x) dx$$

Διακύμανση $V\{X\}$

Ορισμός

Αναμενόμενη τιμή της τυχαίας μεταβλητής $(X-\mu)^2$, δηλαδή των τετραγωνικών αποκλίσεων $(x-\mu^2)$ των τιμών x της μεταβλητής X από την αναμενόμενη τιμή της $\mu = E\{X\}$

Για διακριτές τυχαίες μεταβλητές

Χρησιμοποιώντας K πιθανές τιμές $\{x_1, \dots, x_K\}$:

$$V\{X\} = \sigma^2 = \sum_{k=1}^{K} (x_k - \mu)^2 f_X(x_k)$$

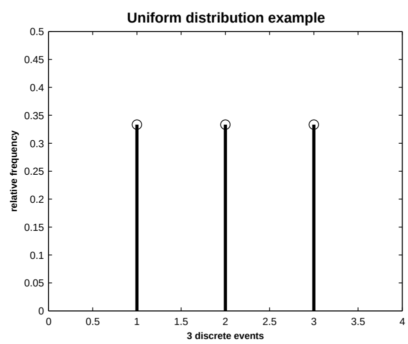
Για συνεχείς τυχαίες μεταβλητές

$$V\{X\} = \sigma^2 = E\{(X-\mu)^2\} = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx$$

Διακριτή Ομοιόμορφη Κατανομή στο Διάστημα $[a, b]$

Συνάρτηση μάζας πιθανότητας

$$f_X(x) = Prob\{X = x\} = \frac{1}{K}, \quad x = 1, \dots, K$$



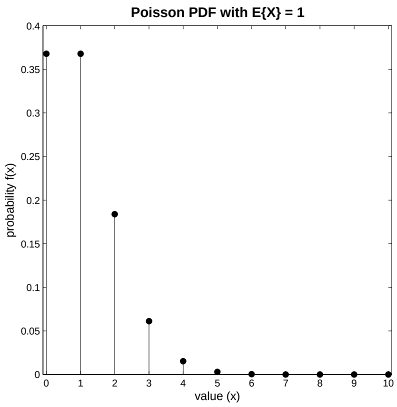
Σημειώσεις

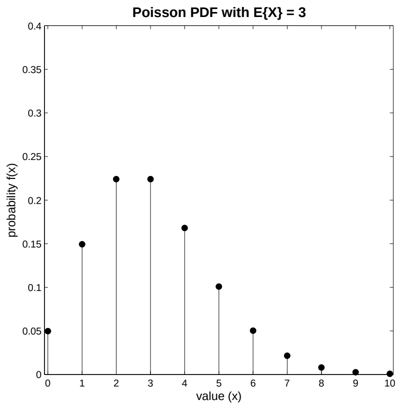
- για μια ομοιόμορφη ΤΜ X: $E\{X\} = \frac{K+1}{2}$, και $V\{X\} = \frac{(K-1)^2}{12}$
- η ομοιλομορφη κατανομή υποδηλώνει ότι οποιαδήποτε (από τις K) τιμή της ΤΜ έχει την ίδια πιθανότητα πραγματοποίησης $1/K$ όπως κάθε άλλη τιμή – έλλειψη γνώσης σχετικά με τα συμβάντα x, εκτός των ορίων $[a, b]$ πραγματοποίησής τους

Η Κατανομή Poisson

Συνάρτηση μάζας πιθανότητας

$$f_X(x) = Prob\{X = x\} = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, \ldots, \infty$$





Σημειώσεις

- αναμενόμενη τιμή $E\{X\} = \mu$: μέσος ρυθμός εμφάνισης (συχνότητα) με μονάδα μέτρησης τον αριθμό εμφάνισης στο μοναδιαίο διάστημα
- για μια Poisson ΤΜ X: $E\{X\} = \mu = V\{X\}$
- χρησιμοποιείται ως μοντέλο κατανομής για τον αριθμό συμβάντων σε μια περίοδο, π.χ., αριθμός ανεμοστρόβιλων σε μια περιοχή μέσα σε ένα διάστημα 100 ετών

Συνεχής Ομοιόμορφη Κατανομή στο Διάστημα $[a, b]$

Όλες οι τιμές της μεταβλητής έχουν την ίδια πιθανότητα πραγματοποίησης

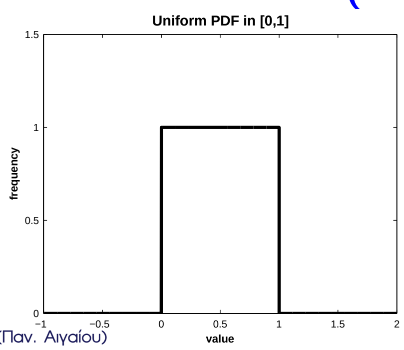
Συνάρτηση πυκνότητας πιθανότητας – ΣΠΠ

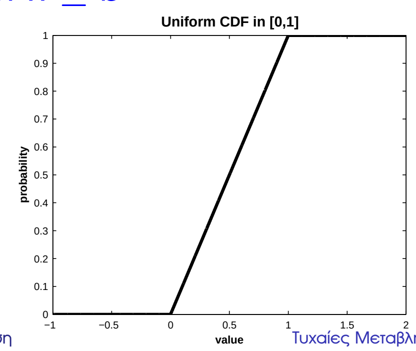
Συνάρτηση $f_X(x)$ με παραμέτρους τα όρια a και b του διαστήματος $[a, b]$. Αναμενόμενη τιμή $E\{X\} = E\{X\} = \frac{b+a}{2}$ και διακύμανση $V\{X\} = \frac{(b-a)^2}{12}$:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{αν } a \leq x \leq b \\ 0 & \text{σε κάθε άλλη περίπτωση} \end{cases}$$

Συνάρτηση αθροιστικής κατανομής – ΣΚ

$$F_X(x) = \begin{cases} 0 & \text{αν } x \leq a \\ \frac{x-a}{b-a} & \text{αν } a \leq x \leq b \\ 1 & \text{αν } x \geq b \end{cases}$$





Εκθετική Κατανομή

Συνεχής μορφή της γεωμετρικής κατανομής, που χρησιμοποιείται για την μοντελοποίηση (τυχαίου) χρόνου αναμονής

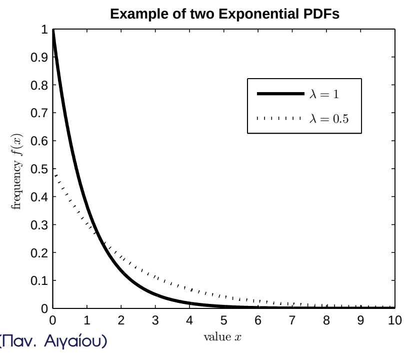
Συνάρτηση πυκνότητας πιθανότητας – ΣΠΠ

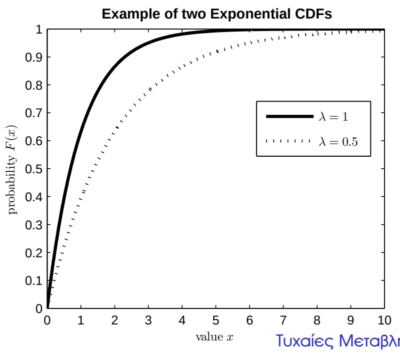
Συνάρτηση $f_X(x)$ με παράμετρο την αναμενόμενη τιμή $E\{X\} = \mu = 1/\lambda$, όπου λ είναι ο ρυθμός συμβάντων στο μοναδιαίο διάστημα. Επίσης, $V\{X\} = \mu^2 = 1/\lambda^2$:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Συνάρτηση αθροιστικής κατανομής – ΣΚ

$$F_X(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$





Τυχαία Δειγματοληψία Από Μια Εκθετική Κατανομή

Διαδικασία

- δημιουργία τυχαίων αριθμών από μια ομοιόμορφη κατανομή στο διάστημα $[0, 1]$:

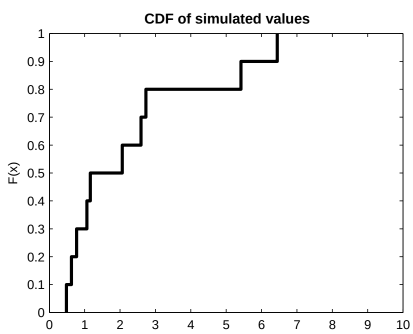
$$\mathbf{p} = [0.2140 \;\; 0.6435 \;\; 0.3200 \;\; 0.9601 \;\; 0.7266 \;\; 0.4120 \;\; 0.7446 \;\; 0.2679 \;\; 0.4399 \;\; 0.9334]$$

- χρησιμοποίηση της συνάρτησης ποσοστημορίων, δηλαδή λύση της εξίσωσης αθροιστικής κατανομής $p = F_X(x)$, εδώ εκθετικής με $\lambda = 1/2$, ως προς το x:

$$p = 1 - e^{-\frac{1}{2}x} \Rightarrow ln(e^{-\frac{1}{2}x}) = ln(1-p) \Rightarrow x = -2\ln(1-p)$$

- για τη δημιουργία προσομοιωμένων τιμών x της μεταβλητής X:

$$\mathbf{x} = [0.4816 \;\; 2.0628 \;\; 0.7713 \;\; 6.4428 \;\; 2.5936 \;\; 1.0621 \;\; 2.7298 \;\; 0.6237 \;\; 1.1593 \;\; 5.4181]$$



Κανονική Κατανομή

Συνάρτηση πυκνότητας πιθανότητας – ΣΠΠ

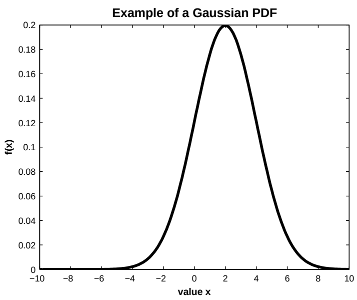
Συνάρτηση $f_X(x)$ με παραμέτρους την αναμενόμενη τιμή $E\{X\} = \mu$ και διακύμανση $V\{X\} = \sigma^2$:

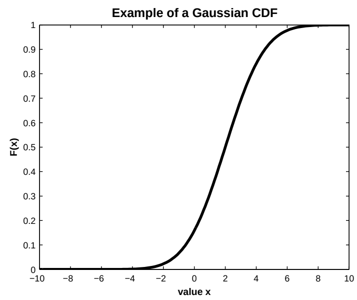
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Συνάρτηση αθροιστικής κατανομής – ΣΚ

Δεν υπάρχει εξίσωση, αλλά υπολογίζεται με αριθμητικές μεθόδους ολοκλήρωσης:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy$$





Κανονική και Τυπική Κανονική Κατανομή

Τυπική και μη τυπική κανονική ΣΠΠ

$$X \sim N(\mu, \sigma^2) \text{ με } f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$Z \sim N(0, 1) \text{ με } f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$Z \sim N(0, 1)$ υποδηλώνει ότι η ΤΜ Z ακολουθεί μια τυπική κανονική ΣΠΠ

Από μη τυπική σε τυπική κανονική ΣΠΠ

κανονικοποίηση των x-δεδομένων, δηλαδή υπολογισμός των z-scores

$$X \sim N(\mu, \sigma^2) \qquad z = \frac{x-\mu}{\sigma} \qquad \text{τώρα } Z \sim N(0, 1)$$

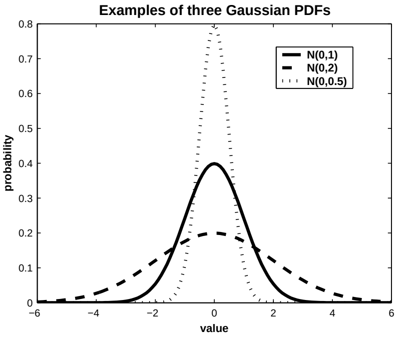
Από τυπική σε μη τυπική κανονική κατανομή

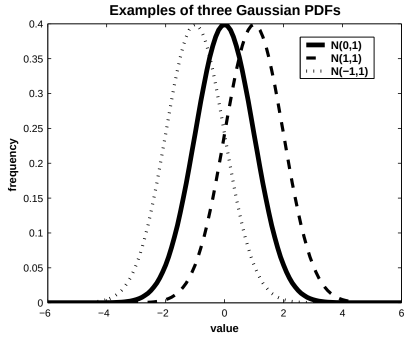
αντίστροφη κανονικοποίηση: πολλαπλασιασμός των z-δεδομένων με την τυπική απόκλιση σ και πρόσθεση της αναμενόμενης τιμής μ:

$$Z \sim N(0, 1) \qquad x = z\sigma + \mu \qquad \text{τώρα } X \sim N(\mu, \sigma^2)$$

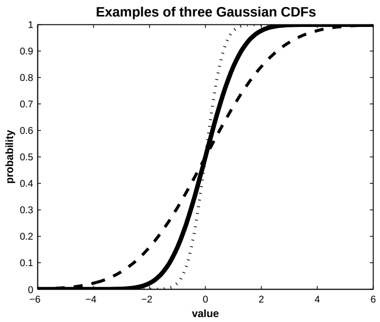
Παραδείγματα

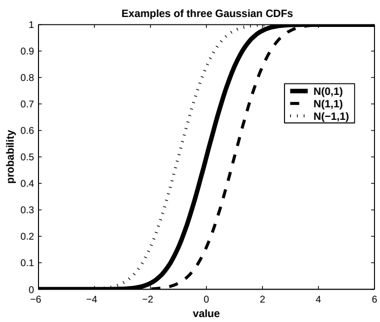
Κανονική ΣΠΠ





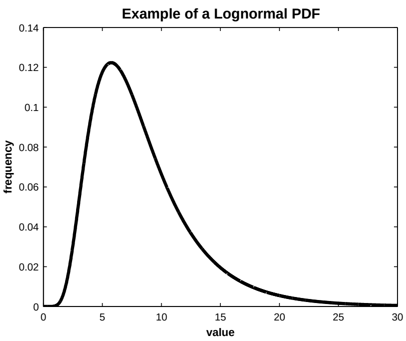
Κανονική ΣΚ

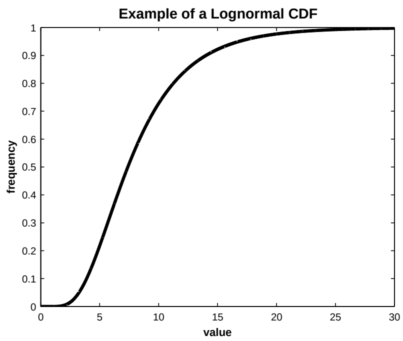




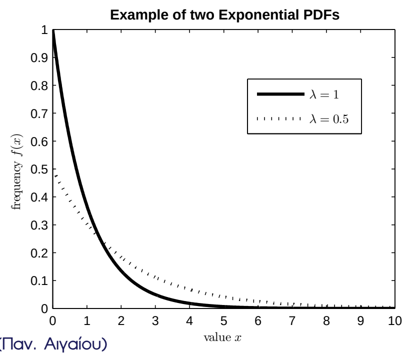
Παραδείγματα (2)

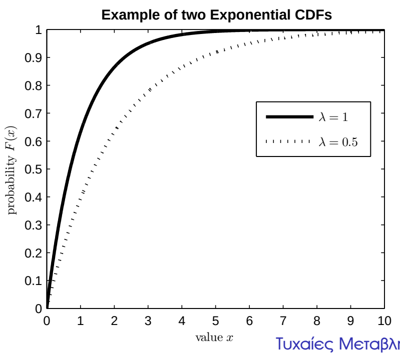
Λογαριθμική ΣΠΠ και ΣΠ





Εκθετική ΣΠΠ και ΣΚ





Θεωρητικές και Δειγματικές ή Εμπειρικές Κατανομές

Προσαρμογή (fitting) θεωρητικής κατανομής σε δεδομένα

- εκτίμηση παραμέτρων της θεωρητικής κατανομής από τα δεδομένα
- χρησιμοποίηση της θεωρητικής κατανομής αντί αυτής των δεδομένων – γενίκευση από την εμπειρική στη θεωρητική κατανομή

Χρησιμότητα θεωρητικών κατανομών

- εξομάλυνση δειγματικής μεταβλητότητας (sampling variation) εξ' αιτίας του μικρού αριθμού δειγμάτων/μετρήσεων
- συμπύκνωση πληροφορίας και παρέκταση (extrapolation) πιθανοτήτων για τιμές που δεν έχουν παρατηρηθεί στο δείγμα

Σημειώσεις

- υπάρχουν πολλά μοντέλα (εξισώσεις) θεωρητικών κατανομών, τα περισσότερα από τα οποία διαθέτουν μία μόνο επικρατούσα τιμή (mode)
- όμως, πολλές μεταβλητές στη φύση δεν είναι δυνατόν να μοντελοποιηθούν με μια απλή παραμετρική κατανομή
- υπάρχουν και πιο γενικευμένα μοντέλα μειγμάτων θεωρητικών κατανομών (με περισσότερες της μίας επικρατούσες τιμές), που παρουσιάζουν όμως μεγαλύτερη δυσκολία στην εκτίμηση των (περισσότερων) παραμέτρων τους.

Πέρα Από τις Μονομεταβλητές Κατανομές

Πολυμεταβλητές κατανομές

- έως το σημείο αυτό, έγινε μια επισκόπηση των κυριότερων θεωρητικών κατανομών, οι οποίες αφορούν στην περιγραφή ενός μόνο χαρακτηριστικού ή μεταβλητής ενός πληθυσμού
- στη χωρική στατιστική, όμως, συχνά υπεισέρχεται η έννοια της πολυμεταβλητής κατανομής, δηλαδή μιας κατανομής που ορίζεται σε ένα πολυδιάστατο (>1) χώρο και περιγράφει την από κοινού κατανομή 2 ή περισσότερων μεταβλητών
- η γενική μορφή μιας διμεταβλητής, για παράδειγμα, συνάρτησης πυκνότητας πιθανότητας (ΣΠΠ) είναι:

$$f_{X_1,X_2}(x_1, x_2; \boldsymbol{\theta})$$

όπου X_1 και X_2 δηλώνουν 2 τυχαίες μεταβλητές, x_1 και x_2 δηλώνουν 2 συγκεκριμένες τιμές (πραγματοποιήσεις) των ΤΜ, και $\boldsymbol{\theta}$ είναι ένα διάνυσμα παραμέτρων, με στοιχεία όπως η αναμενόμενες τιμές και οι τυπικές αποκλίσεις των 2 ΤΜ. Στη συγκεκριμένη περίπτωση, το γράφημα της παραπάνω ΣΠΠ είναι τρισδιάστατο, όπου οι 2 διαστάσεις αντιστοιχούν στις τιμές των 2 μεταβλητών και η 3η διάσταση αντιστοιχεί στις τιμές της πιθανότητας εμφάνισης κάθε **ζεύγους** τιμών $\{x_1, x_2\}$