Ανάλυση Δεδομένων

Δημήτρης Κουγιουμτζής

6 Οκτωβρίου 2020

Περιεχόμενα

1. Εισαγωγή: ορισμοί, δεδομένα, παραδείγματα.
2. Πιθανότητες και Τυχαίες Μεταβλητές, στοιχεία πιθανοτήτων, κατανομές, παράμετροι κατανομής, βασικές κατανομές.
3. Στοιχεία στατιστικής: εκτίμηση παραμέτρων και έλεγχοι υπόθεσης. Υπολογιστικές μέθοδοι (μέθοδοι επαναδειγματοληψίας) στην εκτίμηση παραμέτρων, στατιστικών ελέγχων.
4. Αβεβαιότητα και σφάλμα μέτρησης: συστηματικά και τυχαία σφάλματα, διάδοση σφάλματος.
5. Συσχέτιση και Παλινδρόμηση: συσχέτιση γραμμική και μη-γραμμική, απλή και πολλαπλή παλινδρόμηση, γραμμική και μη-γραμμική παλινδρόμηση. Υπολογιστικές μέθοδοι (μέθοδοι επαναδειγματοληψίας) στην εκτίμηση συσχέτισης και παλινδρόμηση.
6. Μείωση διάστασης για συσχέτιση και παλινδρόμηση: γραμμικές και μη-γραμμικές μέθοδοι.

Βιβλιογραφία μαθήματος (Εύδοξος)

1. *Statistics and Analysis of Scientific Data [electronic resource]*, Bonamente M., Springer ebooks, 2017 (Εύδοξος: 75492930).

2. *Εισαγωγή στην εξόρυξη δεδομένων*, Tan P.-N., Steinbach M., Kumar V., Βερύκιος Βασίλειος (επιλέλεια), Εκδόσεις Τζιόλα & Υιοί, 2018 (2η έκδοση) (Εύδοξος: 77107675).

3. *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining [electronic resource]*, Myatt G.J., Wiley ebooks, 2014 (2nd edition).

Επιπρόσθετη βιβλιογραφία

1. Σημειώσεις 'Ανάλυση δεδομένων', Δ. Κουγιουμτζής, 2020 (είναι μερικώς διαθέσιμες).

2. *Resampling Methods: A Practical Guide to Data Analysis*, Good P.I., Springer ebooks, 2006 (Εύδοξος: 173198)

3. *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*, Wolberg J., Springer ebooks, 2006 (Εύδοξος: 174465)

4. *Computational Statistics Handbook with MATLAB*, Martinez W.L. and Martinez A.R., Chapman and Hall, 3rd edition 2015

5. *Exploratory Data Analysis with MATLAB*, Martinez W.L., Martinez A.R. and Solka J., Chapman and Hall, 3rd edition 2017

6. *Statistical Techniques for Data Analysis*, Taylor J.K. and Cihon C., Chapman and Hall, 2nd edition 2004

data analysis, άλλοι σχετικοί όροι
(data) analytics, big data analytics, data mining, data science

Τι είναι ανάλυση δεδομένων (data analysis);

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making. πηγή: Wikipedia

Τι είναι analytics;

Analytics is the discovery, interpretation, and **communication** of meaningful patterns in data. It also entails applying data patterns towards effective decision making. πηγή: Wikipedia

Τι είναι data analytics;

Data analytics is the science of analyzing raw data in order to make conclusions about that information. πηγή: investopedia

Τι είναι δεδομένα μεγάλης κλίμακας (big data);

"Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. [...] Big data challenges include capturing data, data storage, **data analysis**, search, sharing, transfer, visualization, querying, updating, information privacy and data source. πηγή: Wikipedia

Τι είναι big data analytics;

1. Big data analytics is the often complex process of examining large and varied data sets, or big data, to uncover information – such as hidden patterns, unknown correlations, market trends and customer preferences – that can help organizations make informed business decisions. πηγή: https://searchbusinessanalytics.techtarget.com

Τι είναι big data analytics;

2. Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes. πηγή: https://www.ibm.com

Τι είναι data mining;

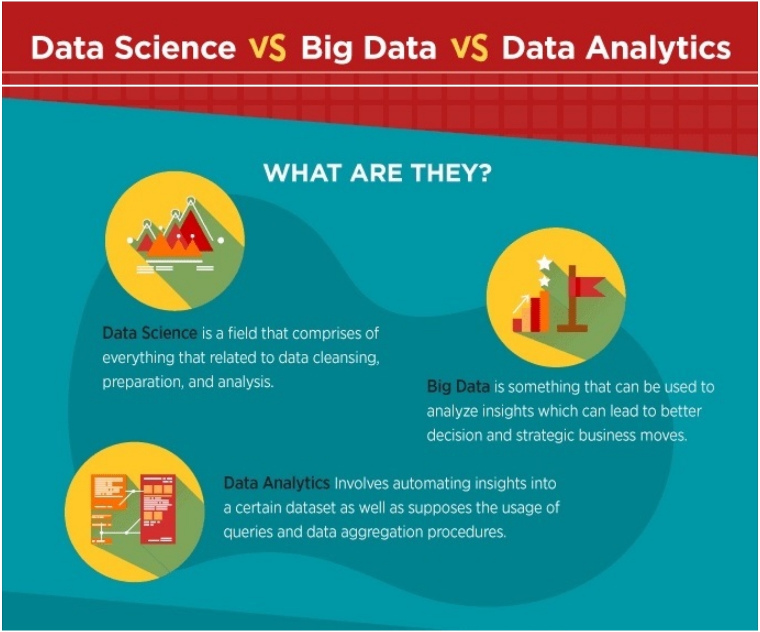
Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, **statistics**, and database systems. Data mining is an interdisciplinary subfield of computer science and **statistics** with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. πηγή: Wikipedia

Τι είναι data science;

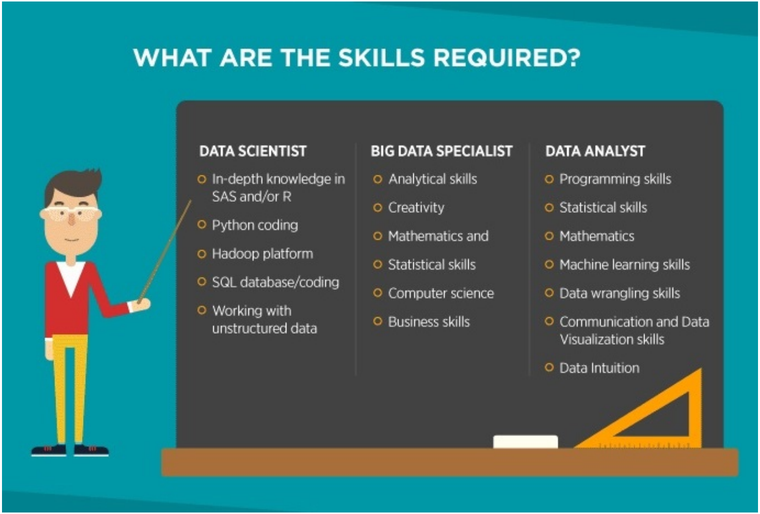
Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems". πηγή: Wikipedia

data analysis, (data) analytics, big data analytics, data mining, data science

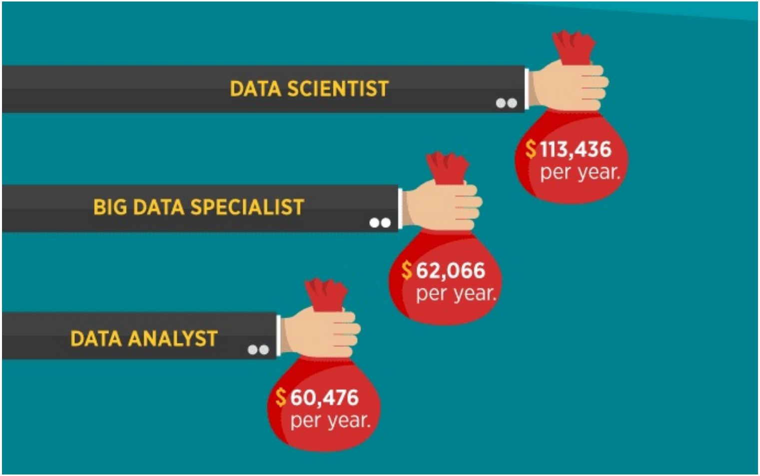
Ποια η διαφορά τους;



πηγή: https://www.simplilearn.com



πηγή: https://www.simplilearn.com



πηγή: https://www.simplilearn.com

Το πλαίσιο της ανάλυσης δεδομένων σε αυτό το μάθημα:

Δεδομένα	
λίγες παρατηρήσεις	✓
πολλές παρατηρήσεις	✓

Ανάλυση	
επιθεώρηση δεδομένων	×
καθαρισμός δεδομένων	×
μετασχηματισμός δεδομένων	✓
μοντελοποίηση δεδομένων	✓
σύνοψη της πληροφορίας σε λίγες παραμέτρους	✓

Προσεγγίσεις ανάλυσης δεδομένων:

- **Αιτιοκρατική** προσέγγιση → μαθηματική ανάλυση
- **Πιθανοκρατική** (στοχαστική) προσέγγιση → πιθανότητες / στατιστική

Τα δεδομένα είναι μετρήσεις μεγέθους/ών ή διαδικασίας

$$\left.\begin{array}{r}\text{Πείραμα}\\ \text{παρακολούθηση}\end{array}\right\}\text{παρατηρήσεις}$$

Δε μπορούμε πάντα να ρωτάμε 'Τι γίνεται αν αλλάξω αυτό;'

Θεωρία Πιθανοτήτων και Στατιστική

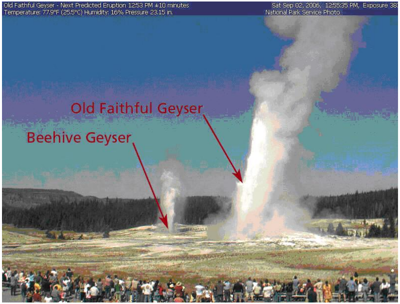
Θεωρία πιθανοτήτων: μελέτη της μεταβλητότητας του αποτελέσματος ενός πειράματος ή διαδικασίας.

Στατιστική:

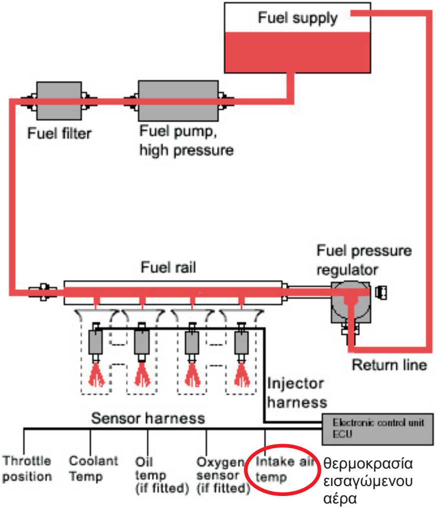
1. *δειγματοληψία*,
2. *περιγραφική στατιστική*,
3. *στατιστική συμπερασματολογία*

Παραδείγματα

1. Η διαφορά τάσης V σε μια αντίσταση
2. Η μέτρηση του μήκους κύματος φασματικής γραμμής με φασματόμετρα
3. Έξαρση θερμοπίδακα



Μεγέθη ενδιαφέροντος:
διάρκεια έξαρσης, χρόνος μεταξύ εξάρσεων



Διάγραμμα συστήματος ψεκασμού καυσίμου

[αντιγραφή στις 9/2012 από τη διεύθυνση http://www.twminduction.com].

Σύστημα ψεκασμού καυσίμου

Τυχαίες μεταβλητές;

θερμοκρασία αέρα, ποσότητα καυσίμου, άνοιγμα βαλβίδας, χρόνος ανάφλεξης, πίεση ψεκασμού

Κατανομή; Μέση τιμή; Διασπορά;

Μοντέλο παλινδρόμησης (θερμοκρασία αέρα από ...)

Θεωρώντας τη χρονική εξέλιξη: ανάλυση χρονοσειράς

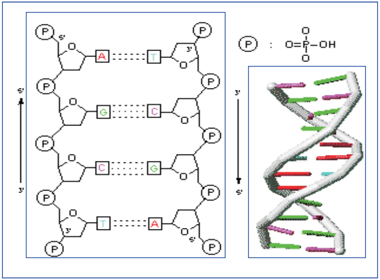
στοχαστική διαδικασία; (αυτοπαλινδρομούμενα μοντέλα)

καθοριστική διαδικασία; (δυναμικό σύστημα, μη-γραμμική ανάλυση χρονοσειρών)

Αντικείμενο μαθήματος 'Ανάλυση χρονοσειρών'

Σειρά DNA

(α)



(β)

Διακριτή τυχαία μεταβλητή $X \in \{A, C, G, T\}$

Ίδια ζητήματα (κατανομή, παλινδρόμηση, συμβολοσειρά)

Ορισμοί

- τυχαία μεταβλητή: συνεχής, διακριτή
- δεδομένα
- πληθυσμός και δείγμα
- παράμετρος και στατιστικό

Ανάλυση δεδομένων / στατιστική:
εκτίμηση παραμέτρων (άγνωστων αλλά σταθερών) του πληθυσμού από τα στατιστικά (γνωστά αλλά μεταβλητά) του δείγματος

Πιθανότητα: σχετική συχνότητα εμφάνισης n_i κάποιας τιμής x_i μιας διακριτής τ.μ. X.

$$\mathrm{P}(x_i) \equiv \mathrm{P}(X = x_i) = \lim_{n \longrightarrow \infty} \frac{n_i}{n}$$

n παρατηρήσεις της X

Υποθέτω στατιστική ομαλότητα.

Για συνεχή τ.μ. X: $\mathrm{P}(x_i) = ?$
Έχει νόημα μόνο $\mathrm{P}(X \in [a, b])$

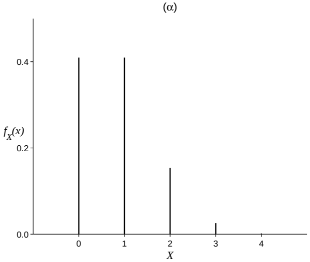
Κατανομή πιθανότητας

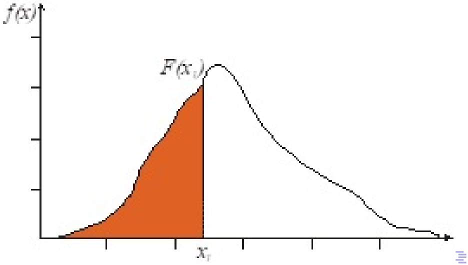
X διακριτή με τιμές $x_1, x_2, \ldots, x_m$:
συνάρτηση μάζας πιθανότητας (pmf) $f_X(x_i) = \mathrm{P}(X = x_i)$
όπου

$$f_X(x_i) \geq 0 \qquad \text{και} \qquad \sum_{i=1}^{m} f_X(x_i) = 1.$$

X συνεχής (π.χ. $X \in \mathbb{R}$):
συνάρτηση πυκνότητας πιθανότητας $f_X(x)$

$$f_X(x) \geq 0 \qquad \text{και} \qquad \int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1.$$





Κατανομή πιθανότητας

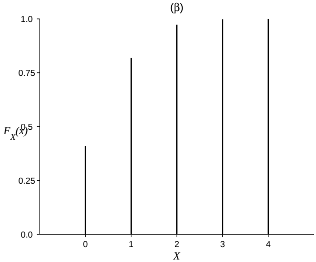
αθροιστική συνάρτηση κατανομής (**cdf**) $F_X(x)$

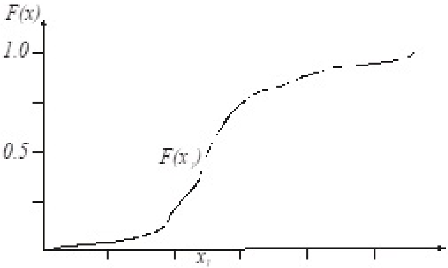
διακριτή:

$$F_X(x_i) = \mathrm{P}(X \leq x_i) = \sum_{x \leq x_i} f_X(x)$$

συνεχή:

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u.$$





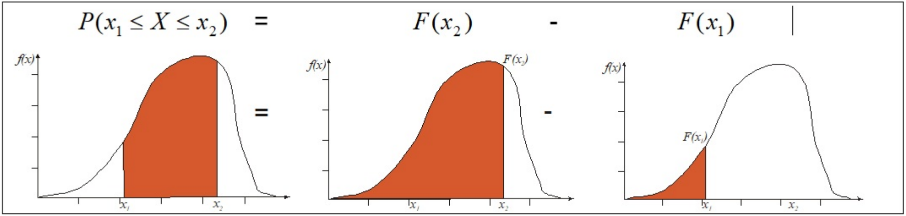
Μετατροπή συνεχής σε διακριτή

Συνεχή σε διακριτή τ.μ. με διαμέριση του πεδίου τιμών

$$\Sigma = \{a \equiv r_0, r_1, \ldots, r_{m-1}, r_m \equiv b\}, \quad \text{όπου} \quad r_0 < r_1 < \ldots < r_m.$$

Αντιστοίχιση τιμών x_i, $i = 1, \ldots, m$, σε κάθε κελί
$[r_{i-1}, r_i)$ $\Rightarrow$ διακριτικοποιημένη τ.μ.X',

$$f_{X'}(x_i) = \mathrm{P}(X' = x_i) = \mathrm{P}(r_{i-1} \leq X \leq r_i) = F_X(r_i) - F_X(r_{i-1}).$$



Από κοινού πιθανότητα δύο τ.μ.

X διακριτή με τιμές $x_1, x_2, \dots, x_n$
Y διακριτή με τιμές $y_1, y_2, \dots, y_m$
από κοινού συνάρτηση μάζας πιθανότητας

$$f_{XY}(x_i, y_i) = \mathrm{P}(X = x_i, Y = y_i)$$

από κοινού (αθροιστική) συνάρτηση κατανομής

$$F_{XY}(x_i, y_i) = \mathrm{P}(X \leq x_i, Y \leq y_i) = \sum_{x \leq x_i} \sum_{y \leq y_i} f_{XY}(x_i, y_i).$$

Από κοινού πιθανότητα δύο τ.μ.

X και Y συνεχείς

από κοινού συνάρτηση πυκνότητας πιθανότητας $f_{XY}(x,y)$

$$f_{XY}(x,y) \geq 0 \qquad \text{και} \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y\mathrm{d}x = 1.$$

από κοινού (αθροιστική) συνάρτηση κατανομής

$$F_{XY}(x,y) = \mathrm{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)\mathrm{d}v\mathrm{d}u.$$

Δύο τ.μ. X και Y (συνεχείς ή διακριτές) είναι **ανεξάρτητες** αν για κάθε δυνατό ζεύγος τιμών τους (x,y) ισχύει

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

Παράμετροι κατανομής τ.μ. - Μέση τιμή, διάμεσος

X διακριτή : $\mu \equiv \mathrm{E}[X] = \sum_{i=1}^{m} x_i f_X(x_i)$

X συνεχής: $\mu \equiv \mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x$

Ιδιότητες:

1. Αν η τ.μ. X παίρνει μόνο μια σταθερή τιμή c είναι $\mathrm{E}[X] = c$.
2. $\mathrm{E}[cX] = c\mathrm{E}[X]$ όπου c σταθερά.
3. X και Y τ.μ.: $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.
4. X και Y ανεξάρτητες τ.μ.: $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Η μέση τιμή έχει τη γραμμική ιδιότητα:
$\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$.
Εκατοστιαία σημεία προσδιορίζονται από την cdf.
διάμεσος $\tilde{\mu}$ είναι το 50-εκατοστιαίο σημείο: $F_X(\tilde{\mu}) = 0.5$.

διασπορά ή **διακύμανση**

$$\sigma^2 \equiv \mathrm{Var}[X] \equiv \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - \mu^2.$$

Ιδιότητες:

1. Αν η τ.μ. X παίρνει μόνο μια σταθερή τιμή c είναι $\mathrm{Var}[c] = 0$.
2. $\mathrm{Var}[X+c] = \mathrm{Var}[X]$ και $\mathrm{Var}[cX] = c^2\mathrm{Var}[X]$, όπου c σταθερά.
3. X και Y ανεξάρτητες τ.μ.: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Η διασπορά δεν έχει τη γραμμική ιδιότητα:
$\mathrm{Var}[aX+bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y]$.

$\mu \equiv \mathrm{E}[X]$ είναι η ροπή πρώτης τάξης και $\mathrm{E}[X^2]$ δεύτερης τάξης
$\sigma^2 \equiv \mathrm{E}[X^2] - \mu^2$ η κεντρική ροπή δεύτερης τάξης.
$\mathrm{E}[X^n]$ ροπή n τάξης
$\mu_n \equiv \mathrm{E}[(x - \mu_X)^n]$ κεντρική ροπή n τάξης.
συντελεστής λοξότητας

$$\lambda = \frac{\mu_3}{\sigma^3} = \mathrm{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right].$$

Για $\lambda = 0$ η κατανομή είναι συμμετρική.
συντελεστή κύρτωσης

$$\kappa = \frac{\mu_4}{\sigma^4} - 3 = \mathrm{E}\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] - 3.$$

Συνδιασπορά και συντελεστής συσχέτισης

συνδιασπορά ή **συνδιακύμανση** δύο τ.μ. X και Y

$$\sigma_{XY} \equiv \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[XY] - \mu_X\mu_Y.$$

Ισχυρή συσχέτιση $\Rightarrow$ μεγάλο σ_{XY}.
συντελεστής συσχέτισης

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Ιδιότητες:

1. $-1 \leq \rho \leq 1$.
2. X και Y ανεξάρτητες $\Rightarrow \rho = 0$, αλλά όχι το αντίθετο.
3. $\rho = -1$ ή $\rho = 1$ αν και μόνο αν $Y = \alpha + \beta X$ για κάποια α και β.

Διωνυμική κατανομή

Επαναλαμβανόμενες δοκιμές Bernoulli: 'επιτυχία' η 'αποτυχία' με την ίδια πιθανότητα p σε κάθε δοκιμή.

X: αριθμός επιτυχίων σε n δοκιμές.
διωνυμική pmf $\mathrm{B}(n, p)$:

$$f_X(x) = \mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

όπου $\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$: *διωνυμικός συντελεστής*.

$$\mu = \mathrm{E}[X] = np \qquad \text{και} \qquad \sigma^2 = \mathrm{Var}[X] = np(1-p).$$

Παράδειγμα για Διωνυμική κατανομή

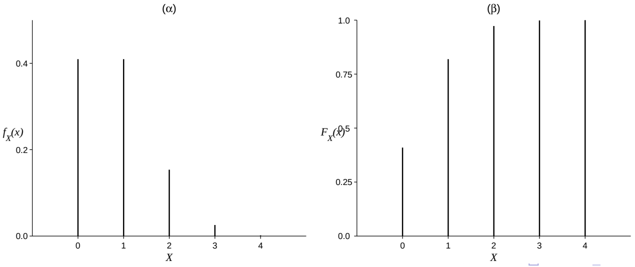
Σε ένα πείραμα αντοχής τάνυσης δοκιμάζουμε 4 βελόνες χαρακτικής σε ένα συγκεκριμένο όριο τάνυσης. Η πιθανότητα να σπάσει η βελόνα σε μια δοκιμή είναι $p = 0.2$. Οι δοκιμές είναι τύπου Bernoulli.

σππ;

$$f_X(0) = \mathrm{P}(X = 0) = \binom{4}{0} 0.2^0 0.8^4 = 0.4096$$

$$f_X(1) = 0.4096 \quad f_X(2) = 0.1536$$

$$f_X(3) = 0.0256 \quad f_X(4) = 0.0016.$$



Παράδειγμα για Διωνυμική κατανομή (συνέχεια)

Η πιθανότητα να σπάσει η βελόνα τουλάχιστον μια φορά

$$\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0) = 1 - 0.4096 = 0.5904$$

η πιθανότητα να σπάσει η βελόνα το πολύ δύο φορές

$$F_X(2) \equiv \mathrm{P}(X \leq 2) = \sum_{x=0}^{2} \mathrm{P}(X = x) = 0.4096 + 0.4096 + 0.1536 = 0.9728$$

Μέση τιμή:
$\mathrm{E}[X] = 4 \cdot 0.2 = 0.8$
δηλαδή στις 4 δοκιμές περίπου μια φορά θα σπάζει η βελόνα.

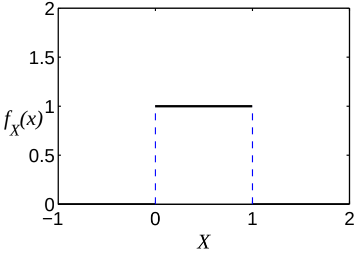
Τυπική απόκλιση:

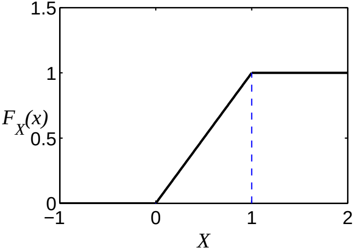
$$\sigma = \sqrt{\mathrm{Var X}} = \sqrt{4 \cdot 0.2 \cdot 0.8} = 0.8.$$

Ομοιόμορφη κατανομή

Η πιό απλή συνεχής κατανομή είναι η ομοιόμορφη κατανομή που ορίζεται σε πεπερασμένο διάστημα $[a, b]$, $X \sim \mathrm{U}[a, b]$

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





$$\mu = \mathrm{E}[X] = \frac{a+b}{2} \qquad \text{και} \qquad \sigma^2 = \mathrm{Var}[X] = \frac{(b-a)^2}{12}.$$

Αντίστροφη της ομοιόμορφης ασκ

Ένα χρήσιμο αποτέλεσμα για δημιουργία τυχαίων αριθμών από οποιαδήποτε γνωστή κατανομή είναι το παρακάτω θεώρημα:

Θεώρημα αντίστροφης ομοιόμορφης ασκ

Αν $X \sim \mathrm{U}[0,1]$ τότε η τ.μ. $Y = F_Y^{-1}(X)$ έχει ασκ $F_Y(y)$.

Παράδειγμα: Τυχαίοι αριθμοί από εκθετική κατανομή.
ασκ εκθετική κατανομής

$$F_Y(y) = 1 - e^{\lambda y}$$

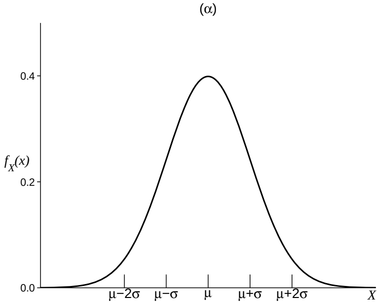
όπου λ η παράμετρος της εκθετικής κατανομής (ίση με μέση τιμή).
Θέτοντας $X \equiv F_Y(y)$, έχουμε $X \sim \mathrm{U}[0,1]$
Για κάθε τιμή x υπολογίζουμε την αντίστοιχη τιμή y από εκθετική κατανομή με παράμετρο λ

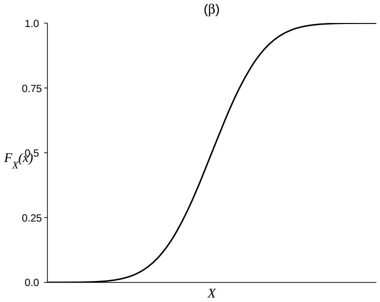
$$y = -\frac{1}{\lambda}\ln(1-x).$$

σππ κανονικής κατανομής

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty,$$

όπου μ μέση τιμή και σ^2 διασπορά
Συμβολισμός: $X \sim \mathrm{N}(\mu, \sigma^2)$.





Τυπική Κανονική κατανομή

τυπική κανονική κατανομή: $Z \sim \mathrm{N}(0, 1)$.

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}} \qquad -\infty < z < \infty.$$

$$\Phi(z) \equiv F_Z(z) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{z} \mathrm{e}^{-\frac{u^2}{2}} \mathrm{d}u \qquad -\infty < z < \infty.$$

Ο μετασχηματισμός της κανονική κατανομής σε τυπική

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad \Longrightarrow \quad Z \equiv \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)$$

επιτρέπει τον υπολογισμό πιθανότητας για X από την $\Phi(z)$

$$\mathrm{P}(a \leq X \leq b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Κανονικότητα

κεντρικό οριακό θέωρημα, ΚΟΘ:
Αν X_i, $i = 1, \ldots, n$ για μεγάλο n
(α) έχουν πεπερασμένη διασπορά και (β) είναι ανεξάρτητες:

$$Y = \sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu_Y, \sigma_Y^2), \quad \mu_Y = \sum_{i=1}^{n} \mu_i, \quad \sigma_Y^2 = \sum_{i=1}^{n} \sigma_i^2$$

Αν X_i έχουν την ίδια κατανομή και μ και σ^2:
$\mu_Y = n\mu$ και $\sigma_Y^2 = n\sigma^2$.
Από το ΚΟΘ ισχύει για το μέσο όρο:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu, \sigma^2/n),$$

δηλαδή $\mu_{\bar{X}} = \mu$ και $\sigma_{\bar{X}}^2 = \sigma^2/n$.

Παράδειγμα για κανονική κατανομή

Το πάχος ενός κυλινδρικού σωλήνα είναι σχεδιασμένο από το εργοστάσιο να είναι μ, αλλά παρατηρείται ότι το πάχος δεν είναι σταθερό σε κάθε παραγόμενο σωλήνα αλλά αποκλίνει από το μ με τυπική απόκλιση $\sigma = 0.1\,\mathrm{mm}$.
Υποθέτουμε λοιπόν ότι το πάχος του κυλινδρικού σωλήνα είναι τυχαία μεταβλητή X που ακολουθεί κανονική κατανομή, δηλαδή $X \sim \mathrm{N}(\mu, 0.1^2)$.

Πιθανότητα η απόκλιση του πάχους να μην είναι μεγαλύτερη από 0.1 mm;

$$\begin{aligned} \mathrm{P}(\mu - 0.1 \leq X \leq \mu + 0.1) &= \mathrm{P}(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) \\ &= 0.8413 - 0.1587 = 0.6826, \end{aligned}$$

... περίπου το 70% των τιμών της X βρίσκονται στο διάστημα $[\mu - \sigma, \mu + \sigma]$.

Παράδειγμα για κανονική κατανομή (συνέχεια)

Όριο σφάλματος ϵ που αντιστοιχεί σε πιθανότητα 0.05;

$$\begin{aligned}
\mathrm{P}(X \leq \mu - \epsilon \text{ ή } X \geq \mu + \epsilon) &= 0.05 \Rightarrow \\
\mathrm{P}(\mu - \epsilon \leq X \leq \mu + \epsilon) &= 0.95 \Rightarrow \\
\Phi\left(\frac{\epsilon}{0.1}\right) - \Phi\left(-\frac{\epsilon}{0.1}\right) &= 0.95 \Rightarrow \\
2\Phi\left(\frac{\epsilon}{0.1}\right) - 1 &= 0.95 \Rightarrow \\
\Phi\left(\frac{\epsilon}{0.1}\right) &= 0.975.
\end{aligned}$$

Στατιστικός πίνακας τυπικής κανονικής κατανομής $\Longrightarrow$
για $\Phi(z) = 0.975$ είναι $z = 1.96$

Με πιθανότητα 0.95 το πάχος του κυλινδρικού σωλήνα δεν αποκλίνει από τη μέση τιμή μ περισσότερο από 0.196 mm.

Ασκήσεις

1. Επιβεβαίωσε τον ορισμό της πιθανότητας ως το όριο της σχετικής συχνότητας για αριθμό επαναλήψεων να τείνει στο άπειρο. Προσομοίωσε τη ρίψη ενός νομίσματος n φορές χρησιμοποιώντας τη γενέτειρα συνάρτηση τυχαίων αριθμών, είτε από ομοιόμορφη διακριτή κατανομή (δίτιμη για 'κορώνα' και 'γράμματα'), ή από ομοιόμορφη συνεχή κατανομή στο διάστημα $[0, 1]$ χρησιμοποιώντας κατώφλι 0.5 (π.χ. αριθμός μικρότερος του 0.5 είναι 'κορώνα' και μεγαλύτερος 'γράμματα'). Επανέλαβε το πείραμα για αυξανόμενα n και υπολόγισε κάθε φορά την αναλογία των 'γραμμάτων' στις n επαναλήψεις. Κάνε την αντίστοιχη γραφική παράσταση της αναλογίας για τα διαφορετικά n.

   Βοήθεια (`matlab`)*:* Για τη δημιουργία των τυχαίων αριθμών χρησιμοποίησε τη συνάρτηση `rand` ή `unidrnd`.

2. Δημιούργησε 1000 τυχαίους αριθμούς από εκθετική κατανομή με παράμετρο $\lambda = 1$ χρησιμοποιώντας την τεχνική που δίνεται στην Παρ. 2.3.2. Κάνε το ιστόγραμμα των τιμών και στο ίδιο σχήμα την καμπύλη της εκθετικής σππ $f_X(x) = \lambda e^{-\lambda x}$.

   Βοήθεια (`matlab`)*:* Το ιστόγραμμα δίνεται με τη συνάρτηση `hist`.

3. Δείξε με προσομοίωση ότι όταν δύο τ.μ. X και Y δεν είναι ανεξάρτητες δεν ισχύει η ιδιότητα $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$. Για να το δείξεις θεώρησε μεγάλο πλήθος τιμών n από X και Y που ακολουθούν τη διμεταβλητή κανονική κατανομή.

   Βοήθεια (`matlab`)*:* Για τον υπολογισμό της διασποράς από n παρατηρήσεις, χρησιμοποίησε τη συνάρτηση `var`. Για να δημιουργήσεις παρατηρήσεις από διμεταβλητή κανονική κατανομή χρησιμοποίησε τη συνάρτηση `mvnrnd`.

4. Ισχύει $\mathrm{E}[1/X] = 1/\mathrm{E}[X]$; Διερεύνησε το υπολογιστικά για X από ομοιόμορφη συνεχή κατανομή στο διάστημα $[1,2]$ υπολογίζοντας τους αντίστοιχους μέσους όρους για αυξανόμενο μέγεθος επαναλήψεων n. Κάνε κατάλληλη γραφική παράσταση για τις δύο μέσες τιμές και τα διαφορετικά n. Τι συμβαίνει αν το διάστημα της ομοιόμορφης κατανομής είναι $[0,1]$ ή $[-1,1]$;

5. Το μήκος X των σιδηροδοκών που παράγονται από μια μηχανή, είναι γνωστό ότι κατανέμεται κανονικά $X \sim \mathrm{N}(4, 0.01)$. Στον ποιοτικό έλεγχο που ακολουθεί αμέσως μετά την παραγωγή απορρίπτονται όσοι σιδηροδοκοί έχουν μήκος λιγότερο από 3.9. Ποια είναι η πιθανότητα μια σιδηροδοκός να καταστραφεί; Που πρέπει να μπει το όριο για να καταστρέφονται το πολύ το 1% των σιδηροδοκών;

   Βοήθεια (`matlab`)*:* Η αθροιστική συνάρτηση κανονικής κατανομής δίνεται με τη συνάρτηση `normcdf`. Η αντίστροφη της δίνεται με τη συνάρτηση `norminv`.

6. Δείξε ότι ισχύει το ΚΟΘ με προσομοίωση. Έστω $n = 100$ τ.μ. από ομοιόμορφη κατανομή στο διάστημα $[0, 1]$ και έστω Y η μέση τιμή τους. Υπολόγισε $N = 10000$ τιμές της Y και σχημάτισε το ιστόγραμμα των τιμών μαζί με την καμπύλη της κανονικής κατανομής.

   Βοήθεια (`matlab`)*:* Το ιστόγραμμα δίνεται με τη συνάρτηση `hist`.