

# Ανάλυση Δεδομένων

Δημήτρης Κουγιουμτζής

21 Οκτωβρίου 2019

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.
4. **Αβεβαιότητα και σφάλμα μέτρησης**: συστηματικά και τυχαία σφάλματα, διάδοση σφάλματος.
5. **Συσχέτιση και Παλινδρόμηση**: συσχέτιση, απλή και πολλαπλή παλινδρόμηση, γραμμική και μη-γραμμική παλινδρόμηση.
6. **Χρονοσειρές**: βασικά χαρακτηριστικά χρονοσειράς, συσχέτιση σε χρονοσειρά.

1. Σημειώσεις 'Ανάλυση δεδομένων', Δ. Κουγιουμτζής, 2019 (δεν είναι ακόμα διαθέσιμες).
2. *Εφαρμοσμένη Στατιστική*, Μπόρα-Σέντα Ε. και Μωυσιάδης Χ., Εκδόσεις Ζήτη, Θεσσαλονίκη 1997 (Εύδοξος: 11028).
3. *Resampling Methods: A Practical Guide to Data Analysis*, Good P.I., Springer, 2006 (Εύδοξος: 173198)
4. *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*, Wolberg J., Springer, 2006 (Εύδοξος: 174465)
5. *Computational Statistics Handbook with MATLAB*, Martinez W.L. and Martinez A.R., Chapman and Hall, 3rd edition 2015
6. *Exploratory Data Analysis with MATLAB*, Martinez W.L., Martinez A.R. and Solka J., Chapman and Hall, 3rd edition 2017
7. *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, Myatt G.J. and Johnson W.P., Wiley-Interscience, 2nd edition, 2014.
8. *Statistical Techniques for Data Analysis*, Taylor J.K. and Cihon C., Chapman and Hall, 2nd edition 2004
9. *Hyperstat*, βιβλίο στο διαδίκτυο (online Book): <http://davidmlane.com/hyperstat>
10. *Concepts and Applications of Inferential Statistics*, Lowry R., βιβλίο στο διαδίκτυο (online book): <http://vassarstats.net/textbook>

**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;

# Data Science VS Big Data VS Data Analytics

## WHAT ARE THEY?



**Data Science** is a field that comprises of everything that related to data cleansing, preparation, and analysis.



**Big Data** is something that can be used to analyze insights which can lead to better decision and strategic business moves.



**Data Analytics** Involves automating insights into a certain dataset as well as supposes the usage of queries and data aggregation procedures.

πηγή: <https://www.simplilearn.com>



## WHAT ARE THE SKILLS REQUIRED?



### DATA SCIENTIST

- In-depth knowledge in SAS and/or R
- Python coding
- Hadoop platform
- SQL database/coding
- Working with unstructured data

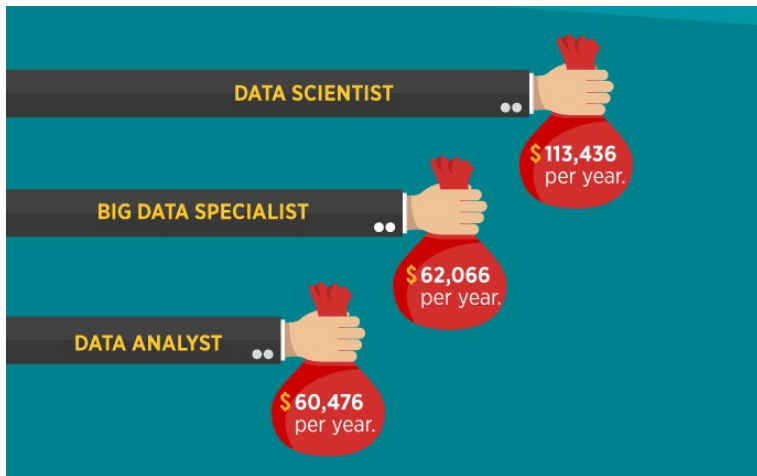
### BIG DATA SPECIALIST

- Analytical skills
- Creativity
- Mathematics and
- Statistical skills
- Computer science
- Business skills

### DATA ANALYST

- Programming skills
- Statistical skills
- Mathematics
- Machine learning skills
- Data wrangling skills
- Communication and Data Visualization skills
- Data Intuition

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

Πείραμα } παρατηρήσεις  
παρακολούθηση }

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



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

**Στατιστική:**

- 1 δειγματοληψία,
- 2 περιγραφική στατιστική,
- 3 στατιστική συμπερασματολογία

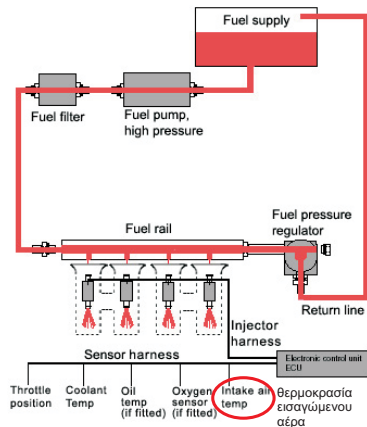
1. Η διαφορά τάσης  $V$  σε μια αντίσταση
2. Η μέτρηση του μήκους κύματος φασματικής γραμμής με φασματόμετρα
3. Έξαρση θερμοπίδακα



Μεγέθη ενδιαφέροντος:

διάρκεια έξαρσης, χρόνος μεταξύ εξάρσεων

# Σύστημα ψεκασμού καυσίμου



Διάγραμμα συστήματος

ψεκασμού καυσίμου (αντιγραφή από τη διεύθυνση <http://www.twminduction.com>).

## Τυχαίες μεταβλητές;

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

Κατανομή; Μέση τιμή; Διασπορά;

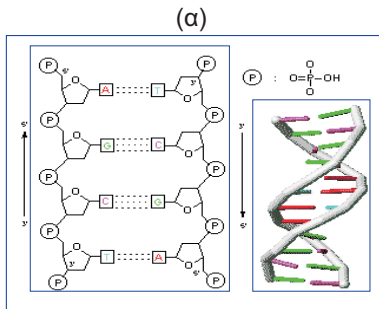
Μοντέλο παλινδρόμησης (θερμοκρασία αέρα από ...)

Θεωρώντας τη χρονική εξέλιξη: **ανάλυση χρονοσειράς**

**στοχαστική διαδικασία;** (αυτοπαλινδρομούμενα μοντέλα)

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

# Σειρά DNA



(β)

```

3721 TTTACCCGGA AACATTGAAATCGGACGGT TTAGTGAAA ATGGAGGATCA AGTTGGGTTT
3781 GGGTCCGTC CGAACGACGA GGAGCTCGTT GGTCACATC TCCGTAACAA AATCGAAGGA
3841 AACACTAGCC GCGACGTTGA AGTAGCCATC AGCGAGGTCA ACATCTGTAG CTACGATCCT
3901 TGGAACTGCG GCTGTAAGTT CGGAATTTTC TGAATTTTAT TTGCAAGTAA TCGATTTAGG
3961 TTTTGTATT TAGGGTTTT TTTGTTTTG AACAG TCCAG TCAAGTACA AATCGAGAGA
4021 TCGATGTGG TACTTCTCT CTCGTAGAGA AAACAACAAA GGAATCGAC ACAGCAGGAC
4081 AACGGTTCCT GGTAAATGGA ACGTTACC GG ABAATCTGTT GAGGTCAAG ACCAGTGGGG
4141 ATTTTGATG GAGGGCTTC GTGGTAAGT TGTGCATAAA AGGGTTTTGG TGTTCCTCGA
4201 TGGAAAGATC CCTGACAAA CCAAATCTGA TTGGGTATC CACGAGTCC ACTACGACCT
4261 CTACAGAAA CATCAG GTTT CTCTTATTC ATATATAT ATATATAT ATGTGGATAT
4321 ATATATATG GGTTCCTGCT GATTCATAG TAGAATTTGA GTTATGCAA TTAGAACTA
4381 TGTATGTAA CTCTATTAG GTTCCAGCAC TATTTAGCG TTAGCTTACT CTCACCAATG
4441 TTTTACTAG ATGAACAT TATGCTTACT CGGAAATTT TACAG AGGAC ATATGTCATC
4501 TGCAGACTG AGTACAAGGG TGATGATGG GACATCTAT CTGCTTATCG AATAGATCCC
4561 ACTCCCGCTT TTGTCCTCAA TATGACTAG ATGCGAGGTT CTGTG GTGAG TCTTTTCCCA
4621 TATACACTTA GCITTTGAGT GGCAGATCAA AAAAGAGCTT GTGTCTACTG ATTTGATGTT
4681 TTCTAAACT GTTGATCGT TTAC GTCAA CCAATCACGT CAACGAAAT CAGGATCTTA
4741 CAACACTTAC TCTGAGTAT ATTCAGCAA TCAATGCCAG CAGTTAAATG AAAACTCTAA
4801 CATTATGCA CAGCAACCAC TTCAAGGATC ATTCAACCCT CTCTTGAAT ATGATTTTGC
4861 AAATCAGCGG GGTCAAGTGG TGATGACTA TATCAGCTC CAACGACAAG TTCCTTACTT
4921 GGCACCTTAT GAAAATGAGT CGGAGATGAT TTGGAAGCT GTGATTTGAAG AAAATTTGA
4981 GTTTTGGTA GATGAAAGGA CATCATGCA ACAGCATTAC AGTGAATCACC GGCCCAAAA
5041 ACGTGTGCT GGGGTTTTGC CTGATGATAG CAGTGTACT GAAACTGGAT CAATGTAAG
5101 CTTTTTTTAC TCATATATA TCACAACCTA TATCGCTTCT ATATCTACA CGCTGAATTT
5161 TGGCTTTTAA CAG ATTTTCG AAGACACTC GAGTCCACT GATAGTGTG GTAGTTCAGA
5221 TGAACCGGGC CACTCTGTA TAGATGATAT TCATCATTTG ACGATTATGG AGCCTTTGCA
5281 CAATTATAAG GCACAAGAGC AACCAAGCA CGCAGGCAA GAAAG GTT ACACCTTCA
5341 CTGGAACA TCACCTTGT ACGAATCTG AATCAACT TTCTACAAAA GATTAGTCA
5401 AATGACCTCT AAATATGAG CTATGGGCTT GCTTTCAGG T GATAATTCG CAGAAAAGCG
5461 AATCGGAGTG GAAAATGGCT GAAGACTCGA TCAAGATACC TCCATCCACC AACACGGTGA
5521 AGCAGAGCTG GATTTGTTG GCAATGACC ATGGGAAC TACTCAAGAC ATGATCATTT
5581 GTGCTGTTT GTTCACTCC GATCATGTT GGATCATCT TGTGGTTAA GAGTCAAT
5641 CGGATTC TG CTA AAAATTT GTATTTCTA GAATGTGTGT TTTTTTTT TTTTTTTT
    
```



- 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.  $X$ .

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

$n$  παρατηρήσεις της  $X$

Υποθέτω **στατιστική ομαλότητα**.

Για συνεχή τ.μ.  $X$ :  $P(x_i) = ?$

Έχει νόημα μόνο  $P(X \in [a, b])$

$X$  διακριτή με τιμές  $x_1, x_2, \dots, x_m$ :

**συνάρτηση μάζας πιθανότητας** (pmf)  $f_X(x_i) = P(X = x_i)$

όπου

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

$X$  συνεχή (π.χ.  $X \in \mathbb{R}$ ):

**συνάρτηση πυκνότητας πιθανότητας**  $f_X(x)$

$$f_X(x) \geq 0 \quad \text{και} \quad \int_{-\infty}^{\infty} f_X(x) dx = 1.$$

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

διακριτή:

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

συνεχή:

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

# Μετατροπή συνεχής σε διακριτή

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

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

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

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

## Κοινή πιθανότητα δύο τ.μ.

$X$  διακριτή με τιμές  $x_1, x_2, \dots, x_n$

$Y$  διακριτή με τιμές  $y_1, y_2, \dots, y_m$

**κοινή συνάρτηση μάζας πιθανότητας**

$$f_{XY}(x_i, y_j) = P(X = x_i, Y = y_j)$$

**κοινή (αθροιστική) συνάρτηση κατανομής**

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

## Κοινή πιθανότητα δύο τ.μ.

$X$  και  $Y$  συνεχείς

**κοινή συνάρτηση πυκνότητας πιθανότητας**  $f_{XY}(x, y)$

$$f_{XY}(x, y) \geq 0 \quad \text{και} \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dy dx = 1.$$

**κοινή (αθροιστική) συνάρτηση κατανομής**

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

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

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

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

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

Ιδιότητες:

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

Η μέση τιμή έχει τη γραμμική ιδιότητα:

$$E[aX + bY] = aE[X] + bE[Y].$$

Εκατοστιαία σημεία προσδιορίζονται από την cdf.

**διάμεσος**  $\tilde{\mu}$  είναι το 50-εκατοστιαίο σημείο:  $F_X(\tilde{\mu}) = 0.5$ .



## διασπορά ή διακύμανση

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

Ιδιότητες:

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

Η διασπορά δεν έχει τη γραμμική ιδιότητα:

$$\text{Var}[aX + bY] = a^2 \text{Var}[X] + b^2 \text{Var}[Y].$$

$\mu \equiv E[X]$  είναι η ροπή πρώτης τάξης και  $E[X^2]$  δεύτερης τάξης  
 $\sigma^2 \equiv E[X^2] - \mu^2$  η κεντρική ροπή δεύτερης τάξης.

$E[X^n]$  ροπή  $n$  τάξης

$\mu_n \equiv E[(X - \mu_X)^n]$  κεντρική ροπή  $n$  τάξης.

**συντελεστής λοξότητας**

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

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

**συντελεστή κύρτωσης**

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

# Συνδιασπορά και συντελεστής συσχέτισης

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

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

Ισχυρή συσχέτιση  $\Rightarrow$  μεγάλο  $\sigma_{XY}$ .

**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  $\alpha$  και  $\beta$ .

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

$X$ : αριθμός επιτυχιών σε  $n$  δοκιμές.

**διωνυμική pmf**  $B(n, p)$ :

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

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

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

# Παράδειγμα για Διωνυμική κατανομή

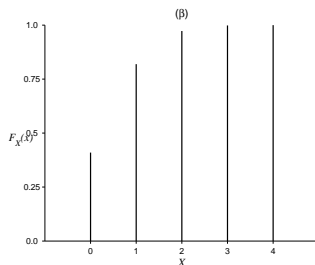
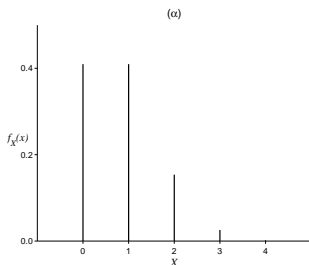
Σε ένα πείραμα αντοχής τάνυσης δοκιμάζουμε 4 βελόνες χαρακτηριστικής σε ένα συγκεκριμένο όριο τάνυσης. Η πιθανότητα να σπάσει η βελόνα σε μια δοκιμή είναι  $p = 0.2$ . Οι δοκιμές είναι τύπου Bernoulli.

σππ;

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

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

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



## Παράδειγμα για Διωνυμική κατανομή (συνέχεια)

Η πιθανότητα να σπάσει η βελόνα τουλάχιστον μια φορά

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

η πιθανότητα να σπάσει η βελόνα το πολύ δύο φορές

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

**Μέση τιμή:**

$$E[X] = 4 \cdot 0.2 = 0.8$$

δηλαδή στις 4 δοκιμές περίπου μια φορά θα σπάζει η βελόνα.

**Τυπική απόκλιση:**

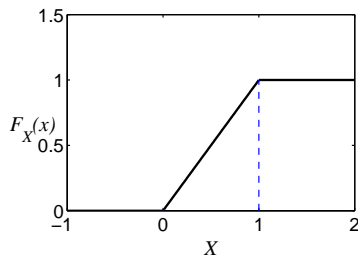
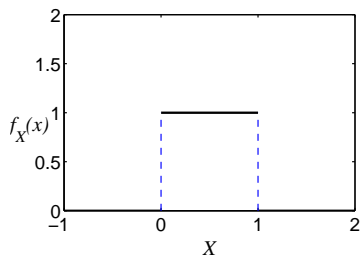
$$\sigma = \sqrt{\text{Var}X} = \sqrt{4 \cdot 0.2 \cdot 0.8} = 0.8.$$

# Ομοιόμορφη κατανομή

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

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

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



$$\mu = E[X] = \frac{a+b}{2} \quad \text{και}$$

$$\sigma^2 = \text{Var}[X] = \frac{(b-a)^2}{12}.$$

# Αντίστροφη της ομοιόμορφης ασκ

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

**Θεώρημα αντίστροφης ομοιόμορφης ασκ**

Αν  $X \sim U[0, 1]$  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  $\lambda$  η παράμετρος της εκθετικής κατανομής (ίση με μέση τιμή).

Θέτοντας  $X \equiv F_Y(y)$ , έχουμε  $X \sim U[0, 1]$

Για κάθε τιμή  $x$  υπολογίζουμε την αντίστοιχη τιμή  $y$  από εκθετική κατανομή με παράμετρο  $\lambda$

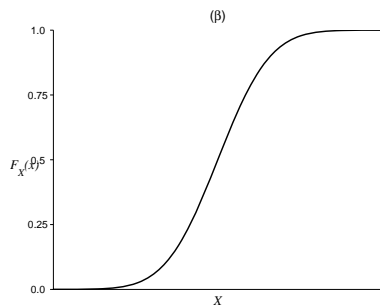
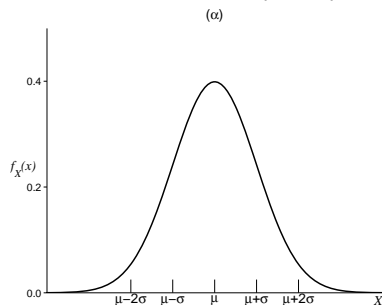
$$y = -\frac{1}{\lambda} \ln(1 - x).$$



## σππ κανονικής κατανομής

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

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



# Τυπική Κανονική κατανομή

**τυπική κανονική κατανομή:**  $Z \sim N(0, 1)$ .

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

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

Ο μετασχηματισμός της κανονική κατανομής σε τυπική

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

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

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

## κεντρικό οριακό θέωρημα, ΚΟΘ:

Αν  $X_i$ ,  $i = 1, \dots, n$  για μεγάλο  $n$

(α) έχουν πεπερασμένη διασπορά και (β) είναι ανεξάρτητες:

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

Αν  $X_i$  έχουν την ίδια κατανομή και  $\mu$  και  $\sigma^2$ :

$\mu_Y = n\mu$  και  $\sigma_Y^2 = n\sigma^2$ .

Από το ΚΟΘ ισχύει για το μέσο όρο:

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

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  $\mu$ , αλλά παρατηρείται ότι το πάχος δεν είναι σταθερό σε κάθε παραγόμενο σωλήνα αλλά αποκλίνει από το  $\mu$  με τυπική απόκλιση  $\sigma = 0.1$  mm.

Υποθέτουμε λοιπόν ότι το πάχος του κυλινδρικού σωλήνα είναι τυχαία μεταβλητή  $X$  που ακολουθεί κανονική κατανομή, δηλαδή  $X \sim N(\mu, 0.1^2)$ .

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

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

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

## Παράδειγμα για κανονική κατανομή (συνέχεια)

Όριο σφάλματος  $\epsilon$  που αντιστοιχεί σε πιθανότητα 0.05;

$$P(X \leq \mu - \epsilon \text{ ή } X \geq \mu + \epsilon) = 0.05 \Rightarrow$$

$$P(\mu - \epsilon \leq X \leq \mu + \epsilon) = 0.95 \Rightarrow$$

$$\Phi\left(\frac{\epsilon}{0.1}\right) - \Phi\left(-\frac{\epsilon}{0.1}\right) = 0.95 \Rightarrow$$

$$2\Phi\left(\frac{\epsilon}{0.1}\right) - 1 = 0.95 \Rightarrow$$

$$\Phi\left(\frac{\epsilon}{0.1}\right) = 0.975.$$

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

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