# Ανάλυση Δεδομένων

Δημήτρης Κουγιουμτζής

7 Μαρτίου 2012

# 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.
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.

# Συγγράμματα


1. *Εφαρμοσμένη Στατιστική*, Μπόρα-Σέντα Ε. και Μωυσιάδης Χ., Εκδόσεις Ζήτη, Θεσσαλονίκη 1997.

2. *Computational Statistics Handbook with MATLAB*, Martinez W.L. and Martinez A.R., Chapman and Hall, 2002

3. *Exploratory Data Analysis with MATLAB*, Martinez W.L. and Martinez A.R., Chapman and Hall, 2005

4. *Statistical Techniques for Data Analysis*, Taylor J.K. and Cihon C., Chapman and Hall, 2004

5. *Making Sense of Data, A Practical Guide to Exploratory Data Analysis and Data Mining*, Myatt G.J., Wiley-Interscience, 2007.

6. *Time Series Analysis, Forecasting and Control*, Box G.E.P., Jenkins G.M. and Reinsel G.C., Prentice Hall, 1994.

7. *Hyperstat*, βιβλίο στο διαδίκτυο (online Book): http://davidmlane.com/hyperstat/

8. *Concepts and Applications of Inferential Statistics*, Lowry R., βιβλίο στο διαδίκτυο (online book): http://faculty.vassar.edu/lowry/webtext.html

# Εισαγωγικά

Σκοπός της **Ανάλυσης δεδομένων**:
Σύνοψη σε λίγες παραμέτρους της πληροφορίας σε ένα σύνολο δεδομένων (που μπορεί να αποτελείται από πολλές παρατηρήσεις).

**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

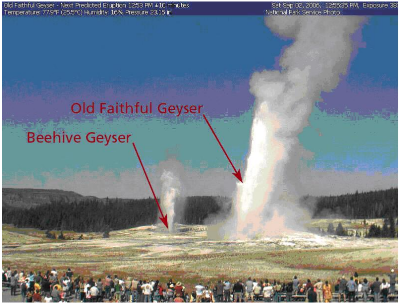
**θεωρία πιθανοτήτων**: μελέτη της μεταβλητότητας του αποτελέσματος ενός πειράματος ή διαδικασίας.

**Στατιστική**:

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

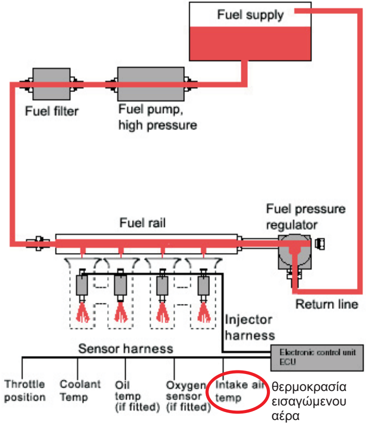
# Παραδείγματα

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

# Σύστημα ψεκασμού καυσίμου




Διάγραμμα συστήματος ψεκασμού καυσίμου (αντιγραφή 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 ...)

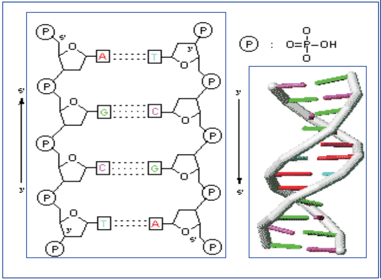
Θεωρώντας τη χρονική εξέλιξη: ανάλυση χρονοσειράς

στοχαστική διαδικασία; (αυτοπαλινδρομούμενα μοντέλα)

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

# Σειρά DNA

(α)




(β)

3721 TTTACCCGGA AACCATTGAA ATCGGACGGT TTAGTGAAA **A TGGAGGATCA AGTTGGGTTT**
**3781 GGGTTCCGTC CGAACGACGA GGAGCTCGTT GGTCACTATC TCCGTAACAA AATCGAAGGA**
**3841 AACACTAGCC GCGACGTTGA AGTAGCCATC AGCGAGGTCA ACATCTGTAG CTACGATCCT**
**3901 TGGAACTTGC GCT**GTAAGTT CCGAATTTTC TGAATTTCAT TTGCAAGTAA TCGATTTAGG
3961 TTTTTGATTT TAGGGTTTTT TTTTGTTTTG AACAG **TCCAG TCAAAGTACA AATCGAGAGA**
**4021 TGCTATGTGG TACTTCTTCT CTCGTAGAGA AAACAACAAA GGGAATCGAC AGAGCAGGAC**
**4081 AACGGTTTCT GGTAAATGGA AGCTTACCGG AGAATCTGTT GAGGTCAAGG ACCAGTGGGG**
**4141 ATTTTGTAGT GAGGGCTTTC GTGGTAAGAT TGGTCATAAA AGGGTTTTGG TGTTCCTCGA**
**4201 TGGAAGATAC CCTGACAAAA CCAAATCTGA TTGGGTTATC CACGAGTTCC ACTACGACCT**
**4261 CTTACCAGAA CATCAG**GTTT TCTTCTATTC ATATATATAT ATATATATAT ATGTGGATAT
4321 ATATATATGT GGTTTCTGCT GATTCATAGT TAGAATTTGA GTTATGCAAA TTAGAAACTA
4381 TGTAATGTAA CTCTATTTAG GTTCAGCAGC TATTTTAGGC TTAGCTTACT CTCACCAATG
4441 TTTTATACTG ATGAACTTAT GTGCTTACCT CCGGAAATTT TACAG **AGGAC ATATGTCATC**
**4501 TGCAGACTTG AGTACAAGGG TGATGATGCG GACATTCTAT CTGCTTATGC AATAGATCCC**
**4561 ACTCCCGCTT TTGTCCCCAA TATGACTAGT AGTGCAGGTT CTGTG**GTGAG TCTTTCTCCA
4621 TATACACTTA GCTTTGAGTA GGCAGATCAA AAAAGAGCTT GTGTCTACTG ATTTGATGTT
4681 TTCCTAAACT GTTGATTCGT TTCAG **GTCAA CCAATCACGT CAACGAAATT CAGGATCTTA**
**4741 CAACACTTAC TCTGAGTATG ATTCAGCAAA TCATGGCCAG CAGTTTAATG AAAACTCTAA**
**4801 CATTATGCAG CAGCAACCAC TTCAAGGATC ATTCAACCCT CTCCTTGAGT ATGATTTTGC**
**4861 AAATCACGGC GGTCAGTGGC TGAGTGACTA TATCGACCTG CAACAGCAAG TTCCTTACTT**
**4921 GGCACCTTAT GAAAATGAGT CGGAGATGAT TTGGAAGCAT GTGATTGAAG AAAATTTTGA**
**4981 GTTTTTGGTA GATGAAAGGA CATCTATGCA ACAGCATTAC AGTGATCACC GGCCCAAAAA**
**5041 ACCTGTGTCT GGGGTTTTGC CTGATGATAG CAGTGATACT GAAACTGGAT CAATG**GTAAG
5101 CTTTTTTTAC TCATATATAA TCACAACCTA TATCGCTTCT ATATCTCACA CGCTGAATTT
5161 TGGCTTTTAA CAG **ATTTTCG AAGACACTTC GAGCTCCACT GATAGTGTTG GTAGTTCAGA**
**5221 TGAACCGGGC CATACTCGTA TAGATGATAT TCCATCATTG AACATTATTG AGCCTTTGCA**
**5281 CAATTATAAG GCACAAGAGC AACCAAAGCA GCAGAGCAAA GAAAAG**GTTT AACACTCTCA
5341 CTGAGAAACA TGACTTTGAT ACGAAATCTG AATCAACATT TCATCAAAAA GATTTAGTCA
5401 AATGACCTCT AAATTATGAG CTATGGGTCT GCTTTCAGG **T GATAAGTTCG CAGAAAAGCG**
**5461 AATGCGAGTG GAAAATGGCT GAAGACTCGA TCAAGATACC TCCATCCACC AACACGGTGA**
**5521 AGCAGAGCTG GATTGTTTTG GAGAATGCAC AGTGGAACTA TCTCAAGAAC ATGATCATTG**
**5581 GTGTCTTGTT GTTCATCTCC GTCATTAGTT GGATCATTCT TGTTGGTTAA** GAGGTCAAAT
5641 CGGATTCTTG CTCAAAATTT GTATTTCTTA GAATGTGTGT TTTTTTTTGT TTTTTTTTCT

# 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

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

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

$X$ συνεχή (π.χ. $X \in \mathrm{R}$):
**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.$X'$,

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

# 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**

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

**κοινή (αθροιστική) συνάρτηση κατανομής**

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

# Κοινή πιθανότητα δύο τ.μ.

$X$ και $Y$ συνεχείς
**κοινή συνάρτηση πυκνότητας πιθανότητας** $f_{XY}(x, y)$

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

**κοινή (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: $F_X(\tilde{\mu}) = 0.5$.

# Παράμετροι κατανομής τ.μ. - Διασπορά

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

# Ροπές μιας τ.μ.

$\mu \equiv \mathrm{E}[X]$ 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 $\sigma_{XY}$.
**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 $\beta$.

# 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

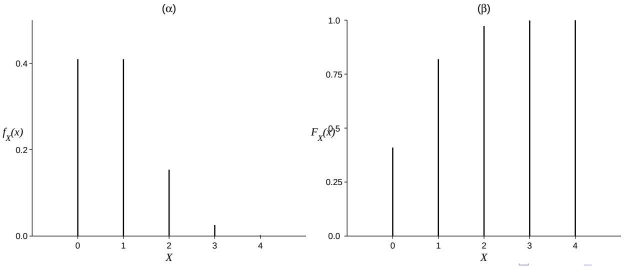
Σε ένα πείραμα αντοχής τάνυσης δοκιμάζουμε 4 βελόνες χαρακτικής σε ένα συγκεκριμένο όριο τάνυσης. Η πιθανότητα να σπάσει η βελόνα σε μια δοκιμή είναι $p = 0.2$. Οι δοκιμές είναι τύπου Bernoulli.

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.

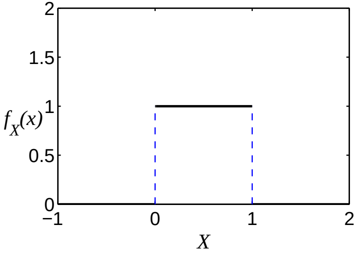
Τυπική απόκλιση:

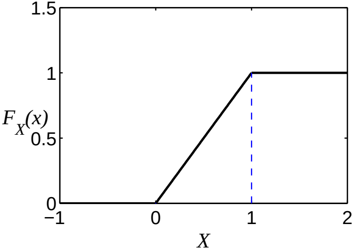
$$\sigma = \sqrt{\mathrm{VarX}} = \sqrt{4 \cdot 0.2 \cdot 0.8} = 0.8.$$

# Ομοιόμορφη κατανομή

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

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




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

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 $\lambda$ η παράμετρος της εκθετικής κατανομής (ίση με μέση τιμή).

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

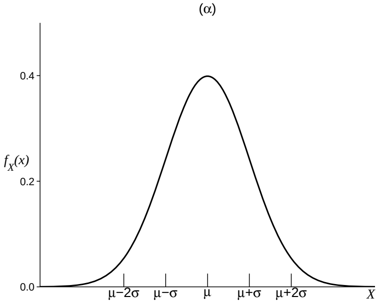
Για κάθε τιμή $x$ υπολογίζουμε την αντίστοιχη τιμή $y$ από εκθετική κατανομή με παράμετρο $\lambda$

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

**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 $\mu$ μέση τιμή και $\sigma^2$ 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 $\Phi(z)$

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

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

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

Αν $X_i$ έχουν την ίδια κατανομή και $\mu$ και $\sigma^2$:
$\mu_Y = n\mu$ και $\sigma_Y^2 = n\sigma^2$.

Από το ΚΟΘ ισχύει για το μέσο όρο:

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

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 $\mu$, αλλά παρατηρείται ότι το πάχος δεν είναι σταθερό σε κάθε παραγόμενο σωλήνα αλλά αποκλίνει από το $\mu$ με τυπική απόκλιση $\sigma = 0.1\,\text{mm}$.
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 $\epsilon$ 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 $\mu$ περισσότερο από 0.196 mm.