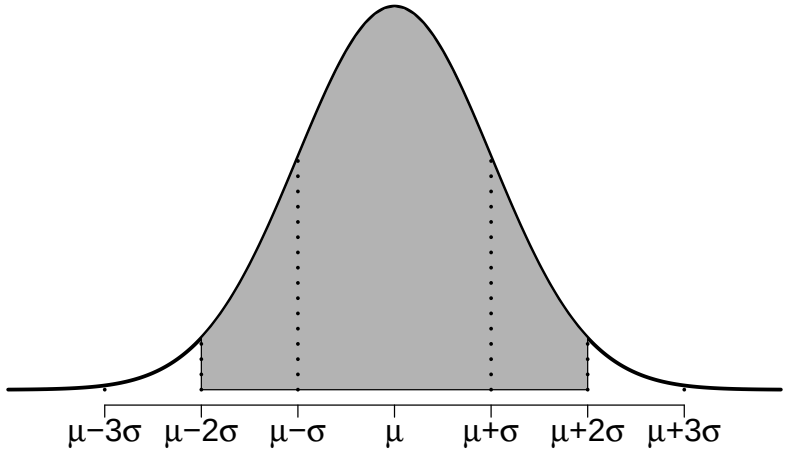
Μάθηση βασισμένη στην πιθανότητες

1 Συνεχή χαρακτηριστικά: Συναρτήσεις πυκνότητας πιθανότητας

Συνεχή χαρακτηριστικά: Συναρτήσεις πυκνότητας πιθανότητας

- Μια **συνάρτηση πυκνότητας πιθανότητας** (PDF) αναπαριστά την κατανομή πιθανότητας ενός συνεχούς χαρακτηριστικού χρησιμοποιώντας μια μαθηματική συνάρτηση, όπως η κανονική κατανομή.

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Μια PDF ορίζει μια καμπύλη πυκνότητας και το σχήμα της καμπύλης καθορίζεται από:
  - τη στατιστική κατανομή που χρησιμοποιείται για τον ορισμό της PDF
  - τις τιμές των παραμέτρων της στατιστικής κατανομής

Πίνακας: Ορισμοί ορισμένων τυπικών κατανομών πιθανότητας.

Κανονική
$x \in \mathbb{R}$
$\mu \in \mathbb{R}$
$\sigma \in \mathbb{R}_{>0}$

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Student-t
$x \in \mathbb{R}$
$\phi \in \mathbb{R}$
$\rho \in \mathbb{R}_{>0}$
$\kappa \in \mathbb{R}_{>0}$
$z = \frac{x-\phi}{\rho}$

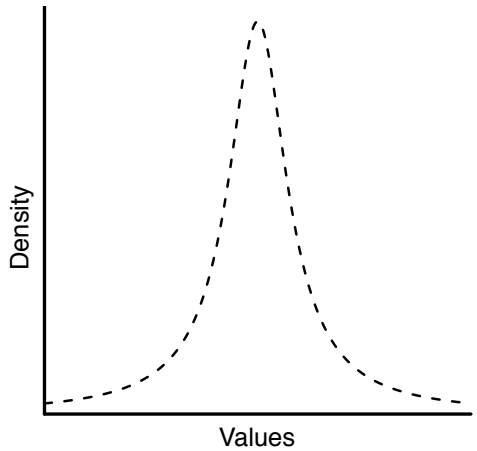
$$\tau(x, \phi, \rho, \kappa) = \frac{\Gamma(\frac{\kappa+1}{2})}{\Gamma(\frac{\kappa}{2}) \times \sqrt{\pi\kappa} \times \rho} \times \left(1 + \left(\frac{1}{\kappa} \times z^2\right)\right)^{-\frac{\kappa+1}{2}}$$

Εκθετική
$x \in \mathbb{R}$
$\lambda \in \mathbb{R}_{>0}$

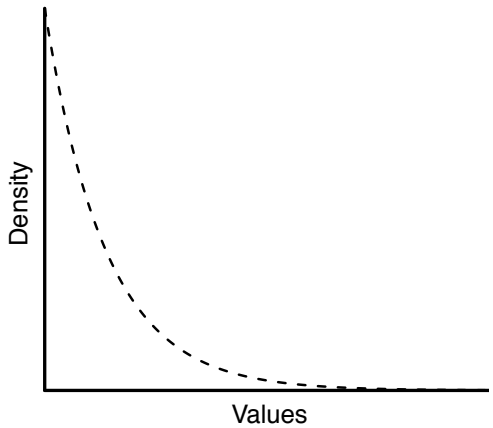
$$E(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{για } x > 0 \\ 0 & \text{διαφορετικά} \end{cases}$$

Μείγμα από n Γκαουσιανές
$x \in \mathbb{R}$
$\{\mu_1, \ldots, \mu_n | \mu_i \in \mathbb{R}\}$
n$\{\sigma_1, \ldots, \sigma_n | \sigma_i \in \mathbb{R}_{>0}\}$
$\{\omega_1, \ldots, \omega_n | \omega_i \in \mathbb{R}_{>0}\}$
$\sum_{i=1}^{n} \omega_i = 0$

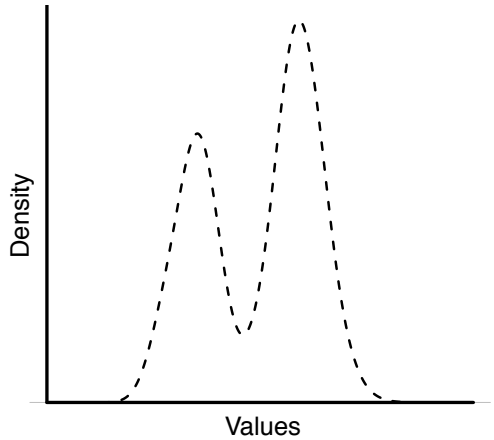
$$N(x, \mu_1, \sigma_1, \omega_1, \ldots, \mu_n, \sigma_n, \omega_n) = \sum_{i=1}^{n} \frac{\omega_i}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$



(α′) Κανονική/Student-t

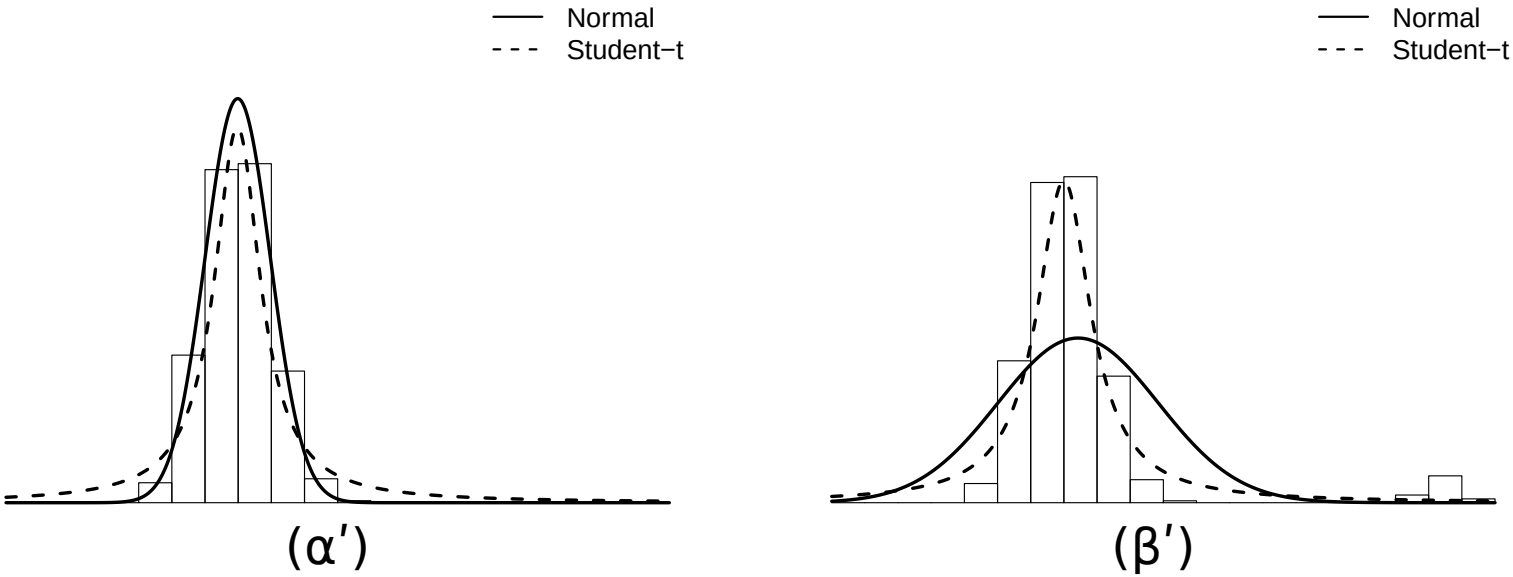


(β′) Εκθετική

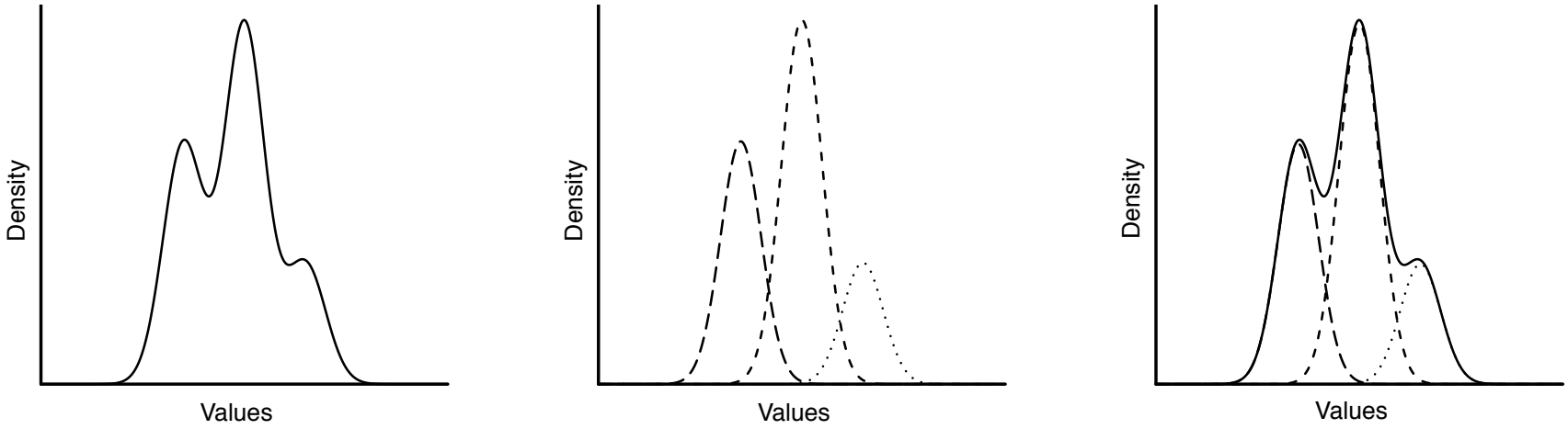


(γ′) Μείγμα Γκαουσιανών

Σχήμα: Γραφήματα ορισμένων γνωστών κατανομών πιθανότητας.

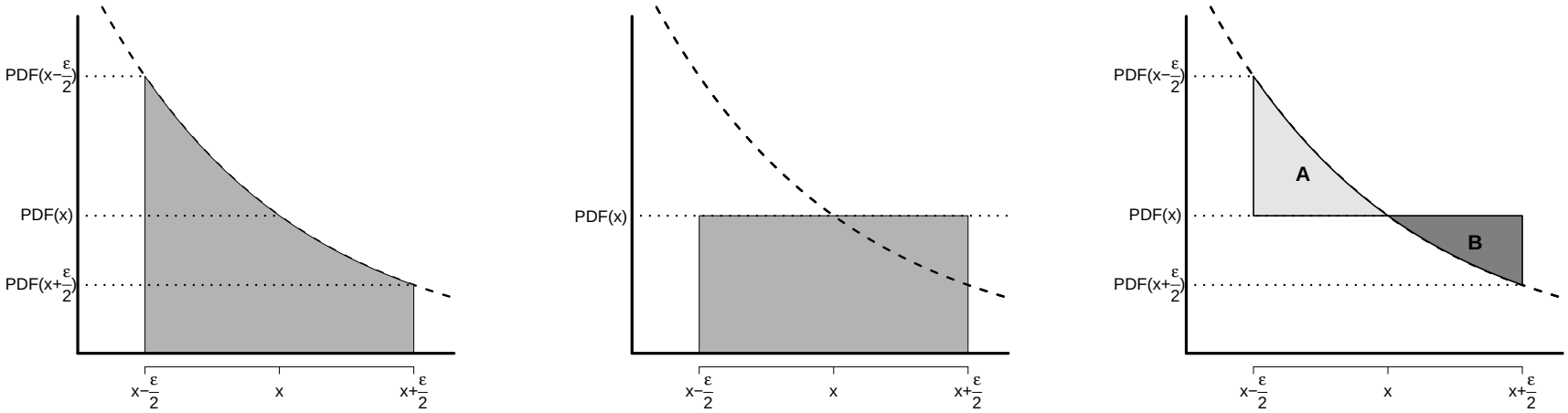


Σχήμα: Απεικόνιση της ανθεκτικότητας της κατανομής student-*t* σε ακραίες τιμές:(a)ιστόγραμμα πυκνότητας ενός συνόλου δεδομένων - καμπυλες πυκνότητας μιας κανονικής και μιας student-*t* κατανομής που έχουν προσαρμοστεί στα δεδομένα(b)ιστόγραμμα πυκνότητας του ίδιου συνόλου δεδομένων με προσθήκη ακραίων τιμών - καμπυλες πυκνότητας μιας κανονικής και μιας student-*t* κατανομής που έχουν προσαρμοστεί στα δεδομένα. Η κατανομή student-*t* επηρεάζεται λιγότερο από εισαγωγή outliers.



Σχήμα: Απεικόνιση του τρόπου με τον οποίο ένα μοντέλο μείγματος Γκαουσιανών αποτελείται από έναν αριθμό κανονικών κατανομών. Η καμπύλη που σχεδιάζεται με συνεχή γραμμή είναι η καμπύλη πυκνότητας του μείγματος Γκαουσιανών, η οποία δημιουργείται με κατάλληλα σταθμισμένο άθροισμα των τριών κανονικών καμπυλών που σχεδιάζονται με διακεκομμένες και τελείες γραμμές.

- Μια PDF είναι μια αφαίρεση πάνω από ένα ιστόγραμμα πυκνότητας και, κατά συνέπεια, αναπαριστά πιθανότητες ως εμβαδό κάτω από την καμπύλη.
- Για να χρησιμοποιήσουμε μια PDF για τον υπολογισμό μιας πιθανότητας πρέπει να σκεφτούμε με όρους εμβαδού κάτω από ένα διάστημα της καμπύλης PDF.
- Μπορούμε να υπολογίσουμε το εμβαδό κάτω από μια PDF είτε αναζητώντας το σε πίνακα πιθανοτήτων είτε χρησιμοποιώντας ολοκλήρωση για να υπολογίσουμε το εμβαδό κάτω από την καμπύλη μέσα στα όρια του διαστήματος.



(α′) (β′) (γ′)

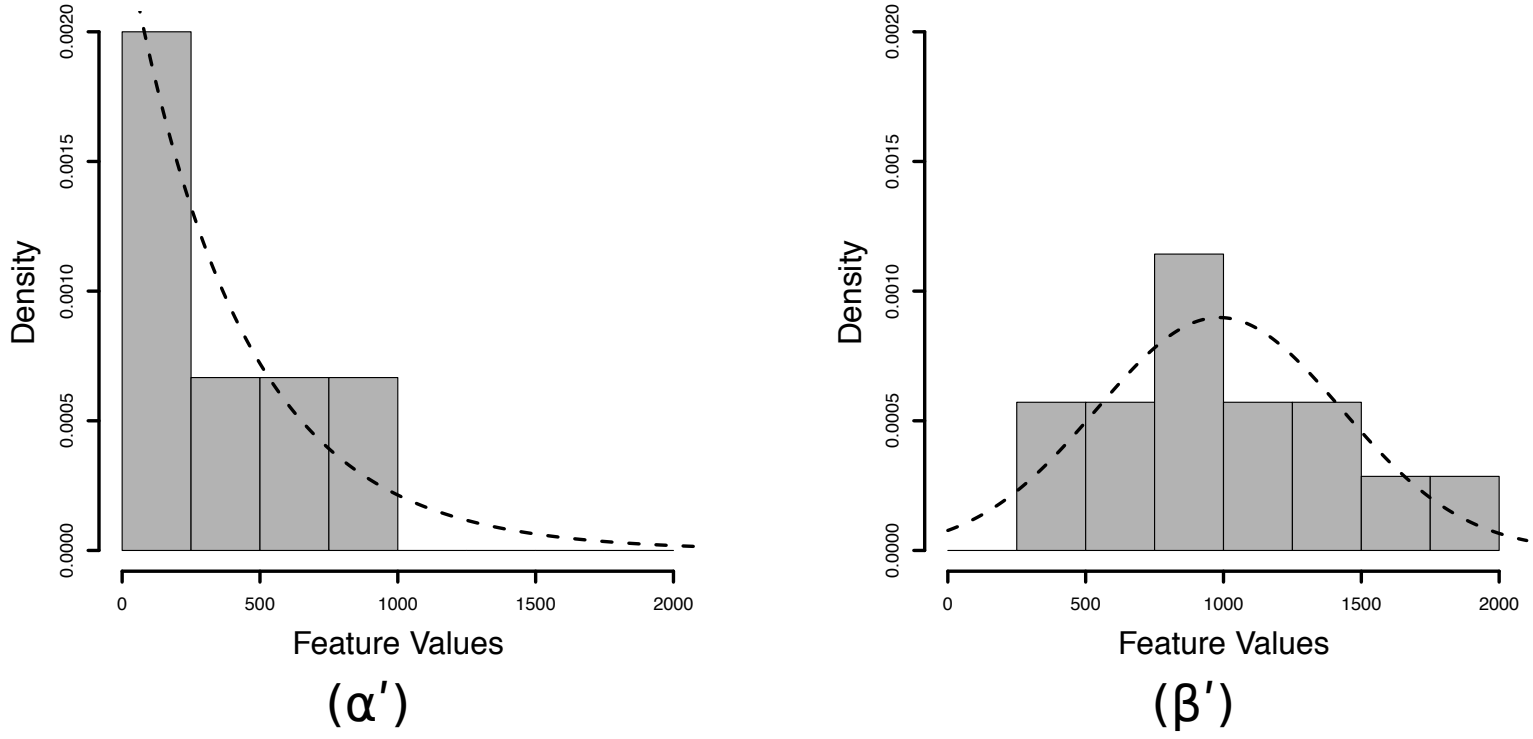
Σχήμα: (α) Το εμβαδό κάτω από μια καμπύλη πυκνότητας μεταξύ των ορίων $x - \frac{\epsilon}{2}$ και $x + \frac{\epsilon}{2}$,(β) η προσέγγιση αυτού του εμβαδού που υπολογίζεται ως $PDF(x) \times \epsilon$, και (γ) το σφάλμα της προσέγγισης είναι ίσο με τη διαφορά μεταξύ της περιοχής A, δηλαδή του εμβαδού κάτω από την καμπύλη που παραλείπεται από την προσέγγιση, και της περιοχής B, δηλαδή του εμβαδού πάνω από την καμπύλη που περιλαμβάνεται λανθασμένα στην προσέγγιση.

- Δεν υπάρχει κάποιος απόλυτος κανόνας για την επιλογή του **μεγέθους διαστήματος** — η απόφαση αυτή λαμβάνεται κατά περίπτωση και εξαρτάται από την ακρίβεια που απαιτείται για την απάντηση σε ένα ερώτημα.
- Για να δείξουμε πώς μπορούν να χρησιμοποιηθούν οι PDF σε μοντέλα Naive Bayes, θα επεκτείνουμε το ερώτημά μας για ανίχνευση απάτης σε αιτήσεις δανείων προσθέτοντας ένα χαρακτηριστικό Account Balance

Πίνακας: Το σύνολο δεδομένων από το πεδίο ανίχνευσης απάτης σε αιτήσεις δανείου με ένα νέο συνεχές περιγραφικό χαρακτηριστικό: Account Balance

ID	Πιστωτικό Ιστορικό	Εγγυητής/ Συναιτών	Κατοικία	Account Balance	Fraud
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears	none	own	1,150.00	false
6	arrears	none	own	928.30	true
7	current	none	own	250.90	false
8	arrears	none	own	806.15	false
9	current	none	rent	1,209.02	false
10	none	none	own	405.72	true
11	current	coapplicant	own	550.00	false
12	current	none	free	223.89	true
13	current	none	rent	103.23	true
14	paid	none	own	758.22	false
15	arrears	none	own	430.79	false
16	current	none	own	675.11	false
17	arrears	coapplicant	rent	1,657.20	false
18	arrears	none	free	1,405.18	false
19	arrears	none	own	760.51	false

- Πρέπει να ορίσουμε δύο PDF για το νέο χαρακτηριστικό Account Balance (AB), με καθεμία να εξαρτάται από διαφορετική τιμή του πεδίου στόχου:
  - $P(AB = X|fr) = PDF_1(AB = X|fr)$
  - $P(AB = X|\neg fr) = PDF_2(AB = X|\neg fr)$
- Σημειώστε ότι αυτές οι δύο PDF δεν είναι απαραίτητο να ορίζονται με την ίδια στατιστική κατανομή.



Σχήμα: Ιστογράμματα, με μέγεθος ομάδος 250 μονάδες, και καμπύλες πυκνότητας για το χαρακτηριστικό Account Balance: (a) τα στιγμιότυπα που δίνουν απάτη σε προσαρμοσμένη εκθετικής κατανομή, (b) τα στιγμιότυπα που δεν δείχνουν απάτη - προσαρμοσμένη κανονική κατανομή.

- Από το σχήμα αυτών των ιστογραμμάτων φαίνεται ότι
  - η κατανομή των τιμών που λαμβάνει το χαρακτηριστικό Account Balance στο σύνολο των παραδειγμάτων όπου το χαρακτηριστικό στόχος Fraud=*'Αληθές'* ακολουθεί εκθετική κατανομή
  - η κατανομή των τιμών που λαμβάνει το χαρακτηριστικό Account Balance στο σύνολο των παραδειγμάτων όπου το χαρακτηριστικό στόχος Fraud=*'Ψευδές'* είναι παρόμοια με κανονική κατανομή.
- Αφού επιλέξουμε τις κατανομές, το επόμενο βήμα είναι να προσαρμόσουμε τις κατανομές στα δεδομένα.

- Για να προσαρμόσουμε την εκθετική κατανομή υπολογίζουμε απλώς τον δειγματικό μέσο όρο, $\bar{x}$, του χαρακτηριστικού Account Balance στο σύνολο των παραδειγμάτων όπου Fraud=*'Αληθές'* και θέτουμε την παράμετρο λ ίση με το αντίστροφο του $\bar{x}$.
- Για να προσαρμόσουμε την κανονική κατανομή στο σύνολο των παραδειγμάτων όπου Fraud=*'Ψευδές'*, υπολογίζουμε απλώς τον δειγματικό μέσο όρο και τη δειγματική τυπική απόκλιση, s, για το χαρακτηριστικό Account Balance σε αυτό το σύνολο και θέτουμε τις παραμέτρους της κανονικής κατανομής ίσες με αυτές τις τιμές.

Πίνακας: Διαμέριση του συνόλου δεδομένων με βάση την τιμή του χαρακτηριστικού στόχου και προσαρμογή των παραμέτρων μιας στατιστικής κατανομής ώστε να μοντελοποιηθεί το χαρακτηριστικό Υπόλοιπο Λογαριασμού σε κάθε διαμέριση.

ID	...	Υπόλοιπο Λογαριασμού	Απάτη
1		56.75	true
4		749.50	true
6		928.30	true
10	...	405.72	true
12		223.89	true
13		103.23	true
$\overline{\text{AB}}$		411.22	
$\lambda = {}^{1}!/_{\overline{\text{AB}}}$		0.0024	

ID	...	Υπόλοιπο Λογαριασμού	Απάτη
2		1 800.11	false
3		1 341.03	false
5		1 150.00	false
7		250.90	false
8		806.15	false
9		1 209.02	false
11		550.00	false
14		758.22	false
15		430.79	false
16		675.11	false
17		1 657.20	false
18		1 405.18	false
19		760.51	false
20		985.41	false
$\overline{\text{AB}}$		984.26	
$sd(\text{AB})$		460.94	