ΤΕΧΝΗΤΗ *ΝΟΗΜΟΣΥΝΗ – ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ*

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ ΤΜΗΜΑ Π.Μ.
ΕΡΓΑΣΤΗΡΙΟ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ Σ.Ε.Π.Μ.

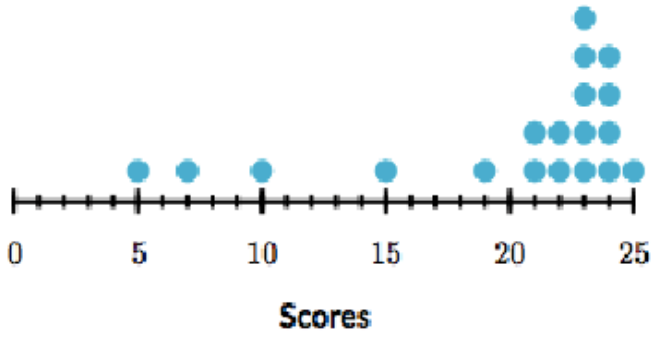
Λάζαρος Ηλιάδης Καθηγητής ΔΠΘ

Email: liliadis@civil.duth.gr

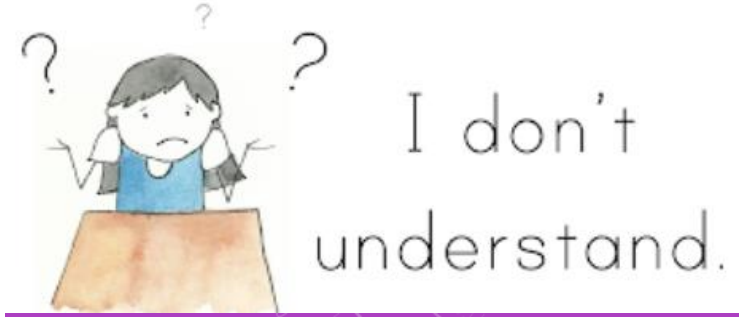
OUTLIERS

OUTLIERS !!

Πιο κάτω βλέπετε τα αποτελέσματα της βαθμολογίας 19 υποψηφίων



CLASSIFICATION !!!!!



Αν σας ρωτήσω κάποιοι θα πείτε ότι υπάρχουν 5 outliers, άλλοι θα πείτε 4, 3, κλπ. Ποιά είναι η αντικειμενική προσέγγιση;
Όταν θα κάνω classification θα πρέπει τα outliers να τα βάλω σε άλλη κατηγορία μακριά από τα υπόλοιπα!

Η μέθοδος λέγεται IQR Μια τιμή είναι outlier αν είναι περισσότερο από 1,5 IQR πάνω από το τρίτο quartile!!!!!!!!!!! ή λιγότερο από το πρώτο quartile!

Δηλαδή είναι κάτω Outlier αν είναι <Q1-1,5 IQR

είναι άνω Outlier αν είναι >Q3+ 1,5IQR

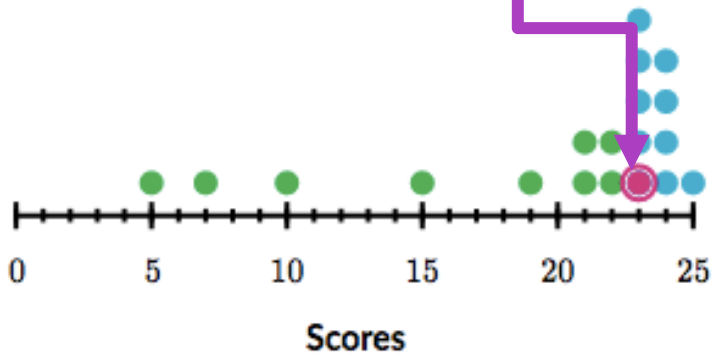
interquartile range (IQR),

9 τιμές

Έστω οι βαθμοί

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Median =23



Το πρώτο Quartile είναι η τιμή που είναι στην μέση των 9 τιμών που είναι κάτω από το Median

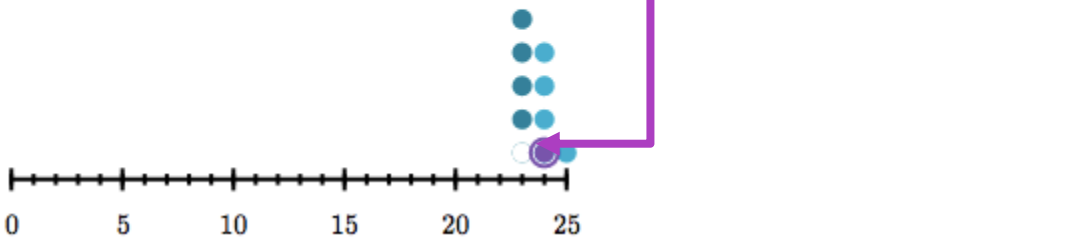
5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

Q1 **Πρώτο quartile =19**

Το τρίτο quartile είναι η μεσαία τιμή των 9 τιμών πάνω από το median

Q3 Τρίτο Quartile το 24

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

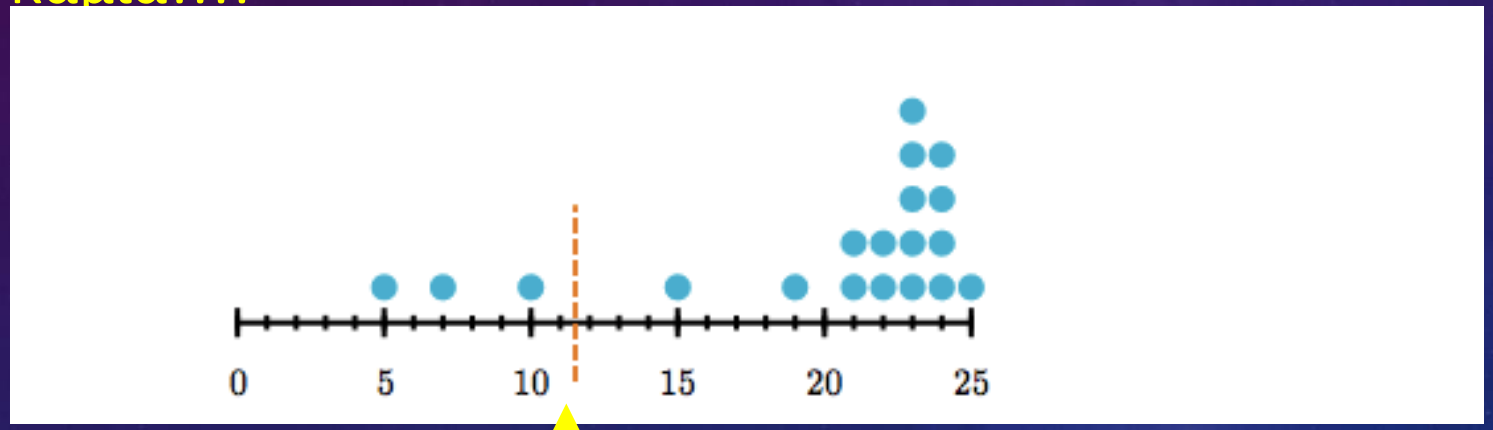


$$IQR = Q_3 - Q_1 = 24 - 19 = 5$$

$$Q_1 - 1.5 \cdot IQR = 19 - 1.5(5) = 19 - 7.5 = 11.5$$

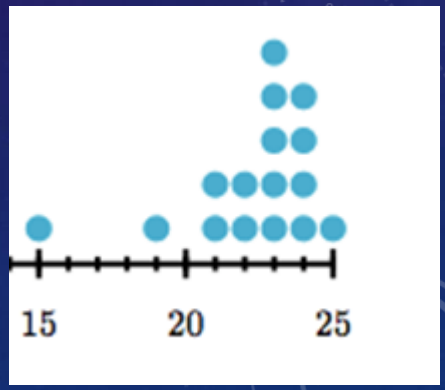
$$Q_3 + 1.5 \cdot IQR = 24 + 1.5(5) = 24 + 7.5 = 31.5$$

Οποιαδήποτε τιμή άνω του Q3 +1,5 IQR = 31,5 είναι ΑΝΩ Outlier. ΔΗΛΑΔΗ Καμία!!!!



Τρεις τιμές κάτω του 11,5 άρα 3 ΚΑΤΩ OUTLIERS!!!!!!!!!!!!

Οι 3 outliers θα είναι η class των Κάτω outliers!!



Classification θα κάνω μόνο με αυτές τις τιμές

Z – SCORE Ή ΑΛΛΙΩΣ STANDARD SCORE

Βασίζεται στην **Μέση Τιμή** (ΜΤ) και στην **Τυπική απ**όκλιση (ΤΑ). Αυτό δημιουργεί προβλήματα, ειδικά όταν αυτές επιρρεάζονται από τα Outliers. Επίσης η μέθοδος συμπεριφέρεται προβληματικά όταν έχουμε μικρά σύνολα δεδομένων. Στην πράξη ποτέ δεν εντοπίζει τα Outliers όταν έχω λιγότερα από 12 δεδομένα.

Το **Modified Z-SCORE** βασίζεται στην **Διάμεσο (Median)** και στο **MAD -Median Absolute Deviation που είναι το Median των απόλυτων αποκλίσεων από τον Median των τιμών.**

$$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|)$$

Είναι ο αριθμός των ΤΑ από τις οποίες μια τιμή δεδομένων είναι Μεγαλύτερη ή Μικρότερη από την Μέση Τιμή.

$$z = \frac{x - \mu}{\sigma}$$

$$z_{\mathrm{Mod}} = \frac{x - \mathrm{Median}}{\mathrm{MAD}}$$

ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ 1

- Η κανονική κατανομή αντιπροσωπεύεται από μια οικογένεια καμπυλών που ορίζεται μοναδικά από δύο παραμέτρους, που είναι η μέση τιμή και η τυπική απόκλιση του πληθυσμού. Οι καμπύλες έχουν πάντα συμμετρικό σχήμα καμπάνας, αλλά ο βαθμός συμπίεσης ή ισοπέδωσης της καμπάνας εξαρτάται από την τυπική απόκλιση του πληθυσμού. Ωστόσο, το απλό γεγονός ότι μια καμπύλη έχει σχήμα καμπάνας δεν σημαίνει ότι αντιπροσωπεύει μια Κανονική κατανομή, επειδή άλλες κατανομές μπορεί να έχουν παρόμοιο είδος σχήματος.

- Πολλά βιολογικά χαρακτηριστικά ακολουθούν μια κανονική κατανομή. Για παράδειγμα, Ύψη ενηλίκων ανδρών και γυναικών, Αρτηριακές πιέσεις σε έναν υγιή πληθυσμό, τυχαία σφάλματα σε πολλούς τύπους εργαστηριακών μετρήσεων και βιοχημικά δεδομένα.

- Όταν ο πληθυσμός από τον οποίο προκύπτουν τα δεδομένα έχει μια κατανομή που είναι περίπου "Κανονική" (ή Gaussian), τότε η τυπική απόκλιση παρέχει μια χρήσιμη βάση για την ερμηνεία των δεδομένων ως προς την πιθανότητα κατανομής τους.

ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ 2

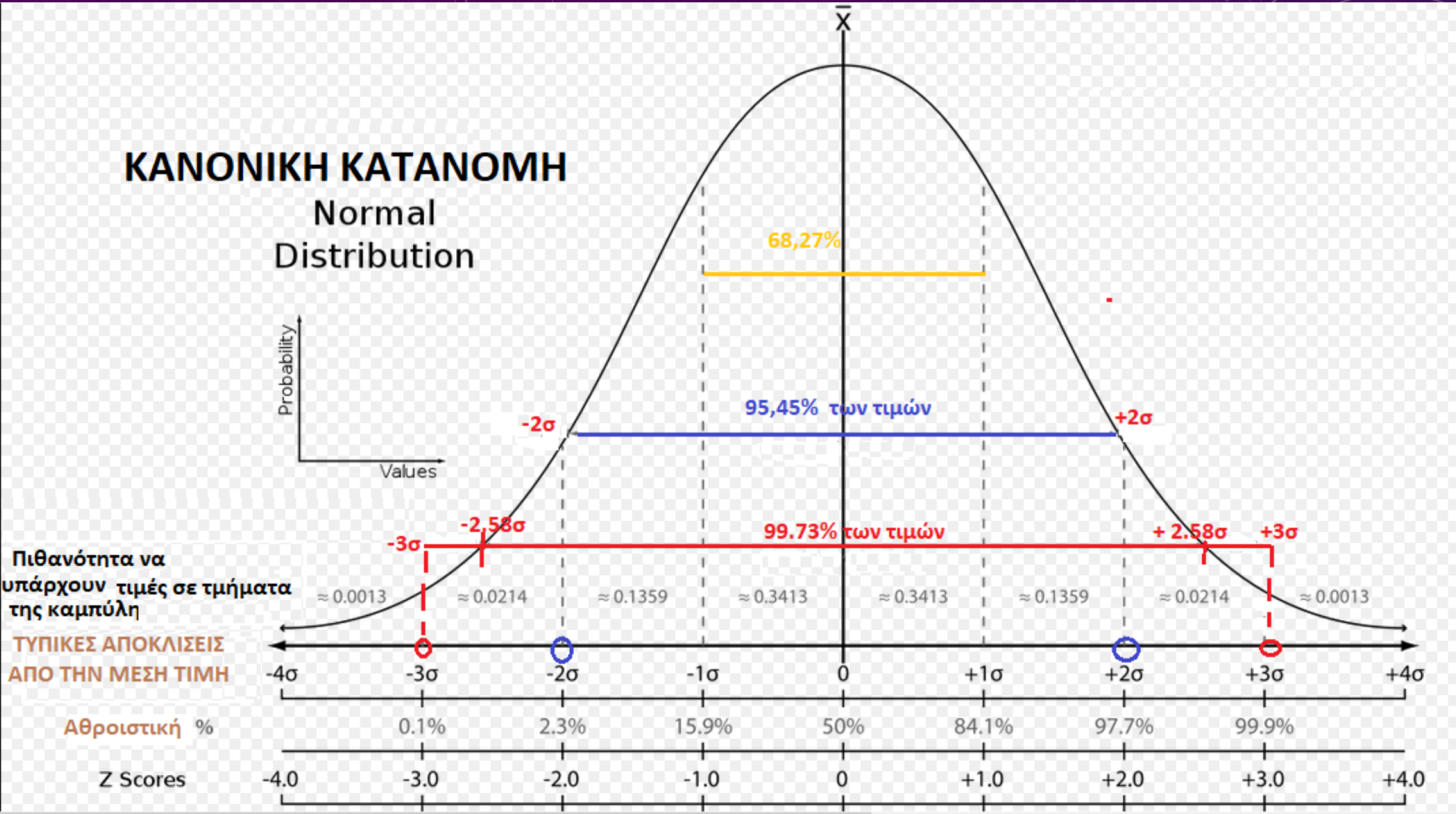
Ο λόγος για τον οποίο η τυπική απόκλιση είναι τόσο χρήσιμο μέτρο της διασποράς των παρατηρήσεων είναι ο εξής:

Εάν οι παρατηρήσεις ακολουθούν μια Κανονική κατανομή, ένα εύρος που καλύπτεται από μία (1σ) τυπική απόκλιση πάνω από τη μέση τιμή και μία τυπική απόκλιση κάτω από αυτήν, περιλαμβάνει περίπου το 68% των παρατηρήσεις; ένα εύρος δύο τυπικών αποκλίσεων (2σ) παραπάνω και δύο παρακάτω περίπου 95% των παρατηρήσεων · και από τρεις τυπικές αποκλίσεις παραπάνω και τρεις παρακάτω () περίπου 99,7% των παρατηρήσεων.

Κατά συνέπεια, εάν γνωρίζουμε τη μέση και τυπική απόκλιση ενός συνόλου παρατηρήσεων, μπορούμε να λάβουμε μερικές χρήσιμες πληροφορίες με απλή αριθμητική. Βάζοντας μία, δύο ή τρεις τυπικές αποκλίσεις πάνω και κάτω από το μέσο όρο, μπορούμε να εκτιμήσουμε τα εύρη που αναμένεται να περιλαμβάνουν περίπου 68%, 95% και 99,7% των παρατηρήσεων.

ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Normal Distribution

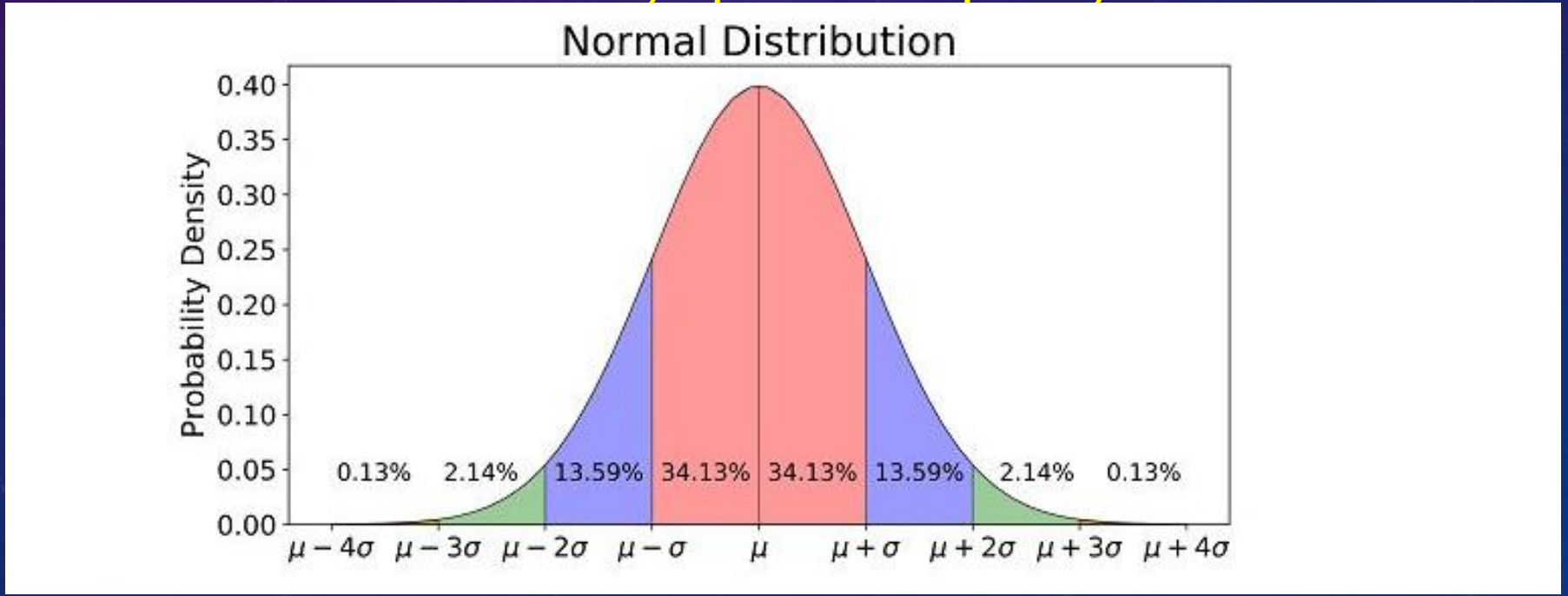


ΑΠΟ ΠΟΙΟ ΣΗΜΕΙΟ ΚΑΙ ΠΕΡΑ ΕΊΝΑΙ OUTLIERS;

Γενικότερα……..

Οποιαδήποτε τιμή 2 ΤΥΠΙΚΕΣ ΑΠΟΚΛΙΣΕΙΣ ΠΑΝΩ από την ΜΤ είναι άνω Outlier ή 3 οποιαδήποτε τιμή 2 ΤΥΠΙΚΕΣ ΑΠΟΚΛΙΣΕΙΣ κάτω από την ΜΤ είναι κάτω **outlier**.

Μπορώ να κάνω τον διαχωρισμό αυτών που είναι > ΜΤ+3σ ή <ΜΤ-3σ ως εξαιρετικά ακραίες



Πως θα ξέρω αν ένα διάνυσμα τιμών ακολουθεί ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ;

Test Kolmogorov – Smirnov

Αν έχω τέλεια ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ τότε το KS Statistic θα είναι ίσο με 0

Επειδή δεν είναι ΑΣΠΡΟ – ΜΑΥΡΟ υπολογίζεται το p-value για να μου δείξει πόσο κοντά είμαι στην ΚΚ.

Το ερώτημα είναι πόσο κοντά είναι το p-value στο 0.05

HTTPS://WWW.SOCSCISTATISTICS.COM/TESTS/KOLMOGOROV/DEFAULT.ASPX

ΑΥΤΌΜΑΤΟΣ ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ P-VALUE ΚΑΙ ΑΠΟΦΑΣΗ !!!

https://www.socscistatistics.com/tests/kolmogorov/default.aspx

Παράδειγμα εκτέλεσης του Υπολογιστή

The Kolmogorov-Smirnov Test of Normality

Success!

Interpreting the Result

The test statistic (D), which you'll see below, provides a measurement of the divergence of your sample distribution from the normal distribution. The higher the value of D, the less probable it is that your data is normally distributed. The *p*-value quantifies this probability, with a low probability indicating that your sample diverges from a normal distribution to an extent unlikely to arise merely by chance. Put simply, high D, low *p*, is evidence that your data *is not* normally distributed.

It's also worth taking a look at the figures provided for skewness and kurtosis. The nearer both these are to zero, the more likely it is that your distribution is normal.

Your Data

```
2
4
5
12
23
65
43
11
21
24
15
6
3
4
2
15
17
18
23
```

Distribution Summary

Count : 19

Mean: 16.47368

Median: 15

Standard Deviation: 15.689233

Skewness: 1.908479

Kurtosis: 4.429

Result: The value of the K-S test statistic (D) is .20579.

The *p*-value is .34819. Your data does *not* differ significantly from that which is normally distributed.

Calculate Reset

ΤΙ ΚΑΝΩ ΑΝ ΘΕΛΩ ΝΑ ΒΡΩ ΣΥΝΔΥΑΣΤΙΚΑ OUTLIERS;

- Π.χ. Θέλω σε ένα διάνυσμα τιμών αέριας ρύπανσης σε μια πόλη να βρώ ΑΚΡΑΙΕΣ ΤΙΜΕΣ CO2 και PM10. (ΔΥΟ ΠΑΡΑΓΟΝΤΕΣ άρα ΔΥΟ ΣΤΗΛΕΣ). Είναι μια μορφή classification!

- Εδώ ξεπερνάω την Στατιστική και δουλεύω με ΜΗ ΕΠΙΤΗΡΟΥΜΕΝΗ ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ (UNSUPERVISED MACHINE LEARNGIG) για να κάνω CLASSIFICATION! Πχ με Fuzzy C-Means Clustering!!

- Αν έχω πάνω από 2 παραμέτρους; Τότε μπορώ να χρησιμοποιήσω INCREMENTAL FUZZY C-MEANS CLUSTERING!!

PRINCIPAL COMPONENT ANALYSIS ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

- https://www.youtube.com/watch?v=FgakZw6K1QQ
- Ένα καταπληκτικό βίντεο που εξηγεί πως λειτουργεί η μέθοδος!
- Μελετήστε το και φτιάξτε ένα power point… που θα μου στείλετε.

PEARSON CORRELATION ANALYSIS

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{(Eq.3)}$$