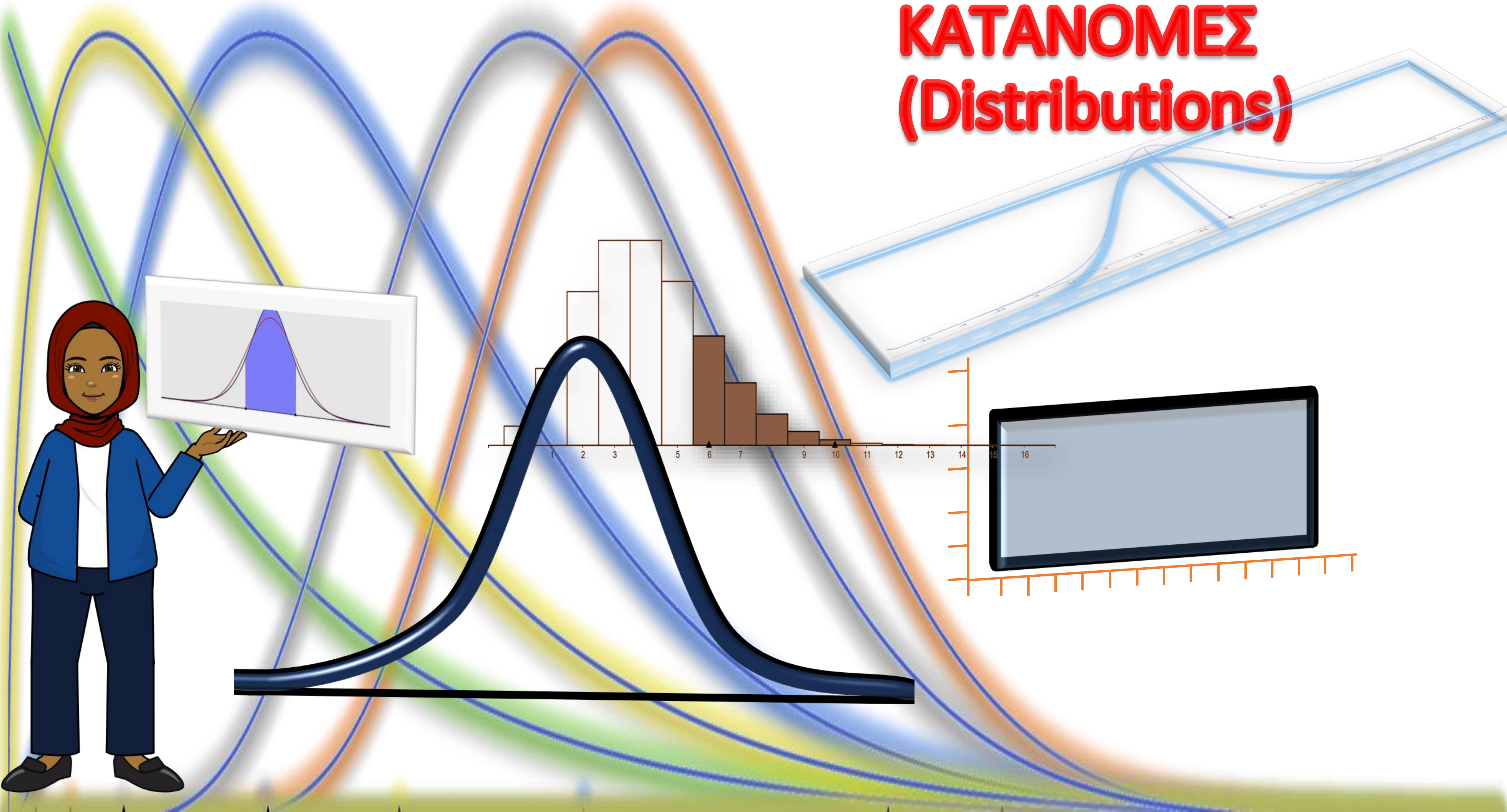
ΚΑΤΑΝΟΜΕΣ (Distributions)



Τι είναι μια κατανομή (πιθανότητων αλλά συχνά ξεχνάμε να το λέμε έτσι)?

Οι πιθανές τιμές που μπορεί να πάρει μια μεταβλητή και πόσο συχνά μπορεί να πάρει τις τιμές αυτές

Τι είναι μια κατανομή?

Παράδειγμα: ποια είναι η κατανομή επιπέδων χοληστερόλης σε ένα αριθμό δειγμάτων

208, 187, 192, 187, 198, 202, 185, 205, 211, 215, 235, 195, 207, 206, 222, 177, 225

1. Τοποθετούμε τα δείγματα κατά σειρά και ομαδοποιούμε αυτά που βρίσκονται μεταξύ κάποιων ορίων→170-179,180-189, 190-199, 200-209, 210-219, 220-229, 230-239
2. Διαιρούμε με τον συνολικό αριθμό δειγμάτων, η πιθανότητα ένα δείγμα να ανήκει μεταξύ κάποιων ορίων π.χ. για την ομάδα 180-189 → 187, 187 (2 δείγματα) / σύνολο δειγμάτων 17 = 2/17= 0.12

Τι είναι μια κατανομή?

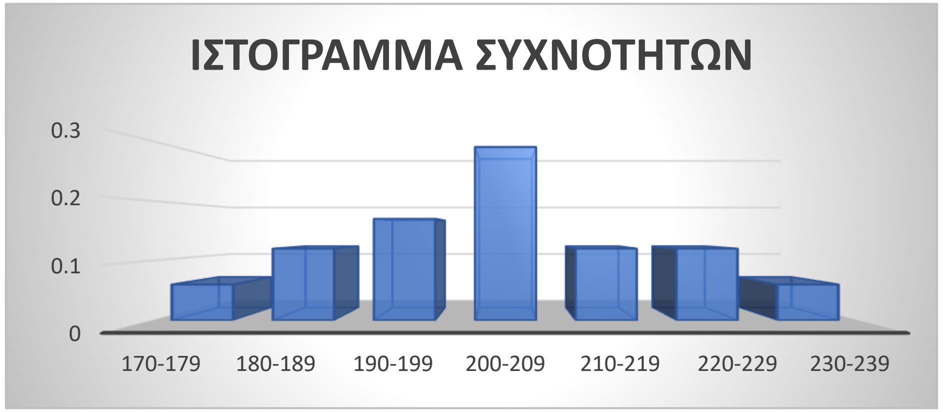
Παράδειγμα: ποια είναι η κατανομή επιπέδων χοληστερόλης χοληστερόλης σε ένα αριθμό δειγμάτων

208, 187, 192, 187, 198, 202, 185, 205, 211, 215, 235, 195, 207, 206, 222, 177, 225

3. Δημιουργούμε ένα **ΠΙΝΑΚΑ ΣΥΧΝΟΤΗΤΩΝ**

170-179	180-189	190-199	200-209	210-219	220-229	230-239
0.06	**0.12**	**0.17**	**0.29**	**0.12**	**0.12**	**0.06**

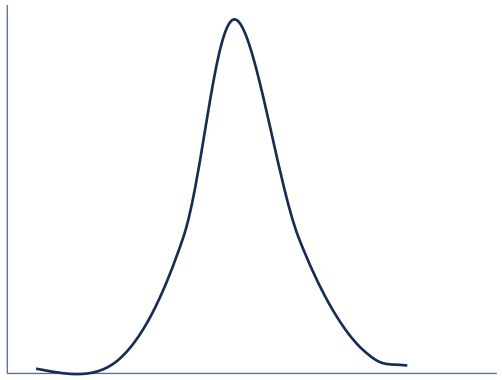
4. Δημιουργούμε το διάγραμμα (συνήθως ένα ιστόγραμμα… για αρχή)



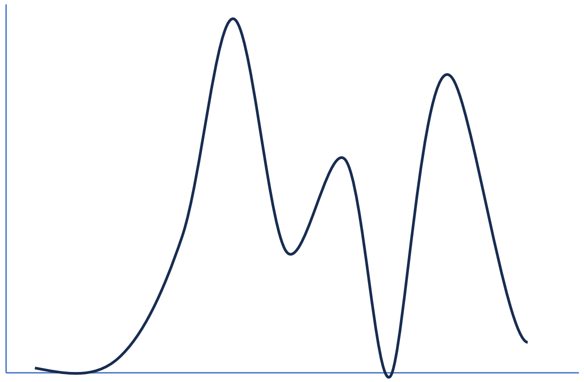
Ιδιότητες μιας κατανομής (πιθανότητων)

Θα μιλήσουμε για τις τρεις βασικές ιδιότητες

Αλλά για να είναι χρήσιμη μια κατανομή πιθανοτήτων πρέπει να ακολουθεί ένα αναγνωρίσιμο μαθηματικό μοντέλο



Μοντέλο Gauss



Μοντέλο otinanai

Αν δεν μπορούμε να την μοντελοποιήσουμε είναι ελάχιστα χρήσιμη
(δεν μπορούμε να τη χρησιμοποιήσουμε…)

Ιδιότητες μιας κατανομής (πιθανότητων)

Με P συμβολίζεται η πιθανότητα

Η γενική πιθανότητα συμβολίζεται με Y → P(Y)

και η συγκεκριμένη πιθανότητα μιας τιμής με y → P(Y=y)

Στο προηγούμενο παράδειγμα πιθανότητα να έχει κάποιος επίπεδα χοληστερόλης 187 είναι P(Y=187) ή για συντομία P(187)=0.12

P(y) → η ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΤΗΤΑΣ

Ιδιότητες μιας κατανομής (πιθανοτήτων)

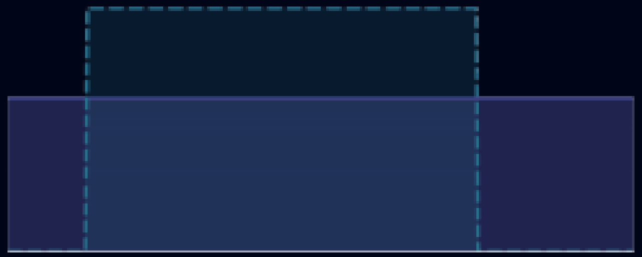
Δυο βασικές ιδιότητες χαρακτηρίζουν τις κατανομές

Μέση τιμή και τυπική απόκλιση

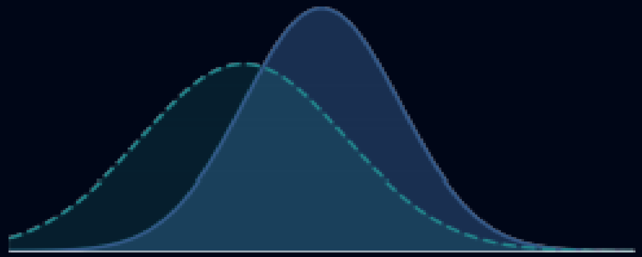
Αλλά πρωταρχικά το σχήμα καθορίζει τη μαθηματική σχέση

Υπάρχουν πολλές κατανομές

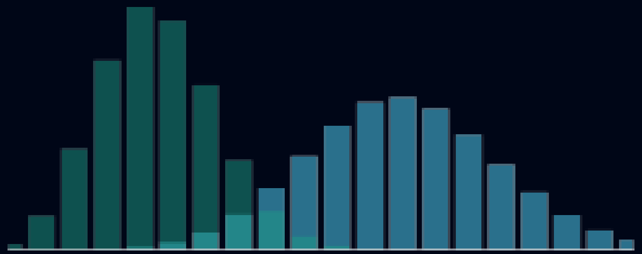
Common Distributions



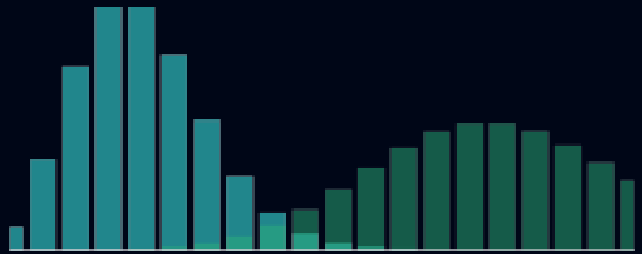
Uniform
Rolling a dice



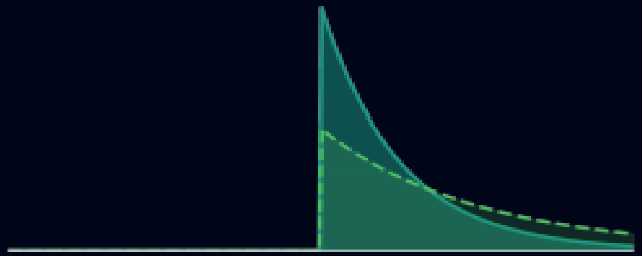
Normal
Central Limit Theorem



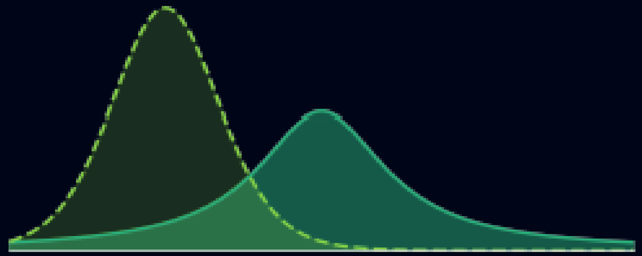
Binomial
Flipping a coin



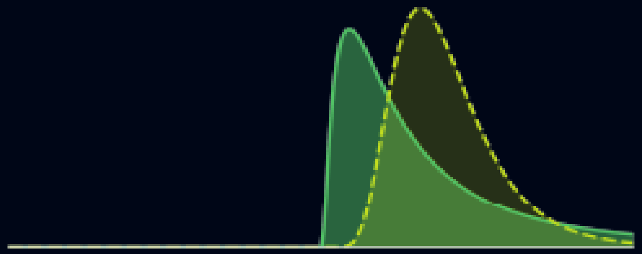
Poisson
Calls in a call center



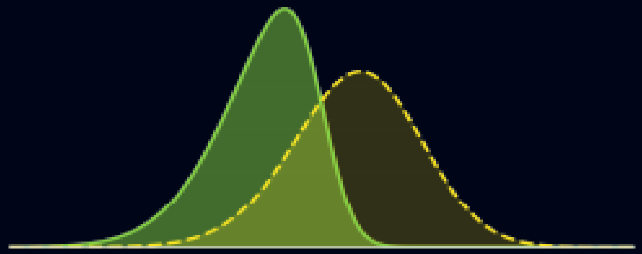
Exponential
Product failure



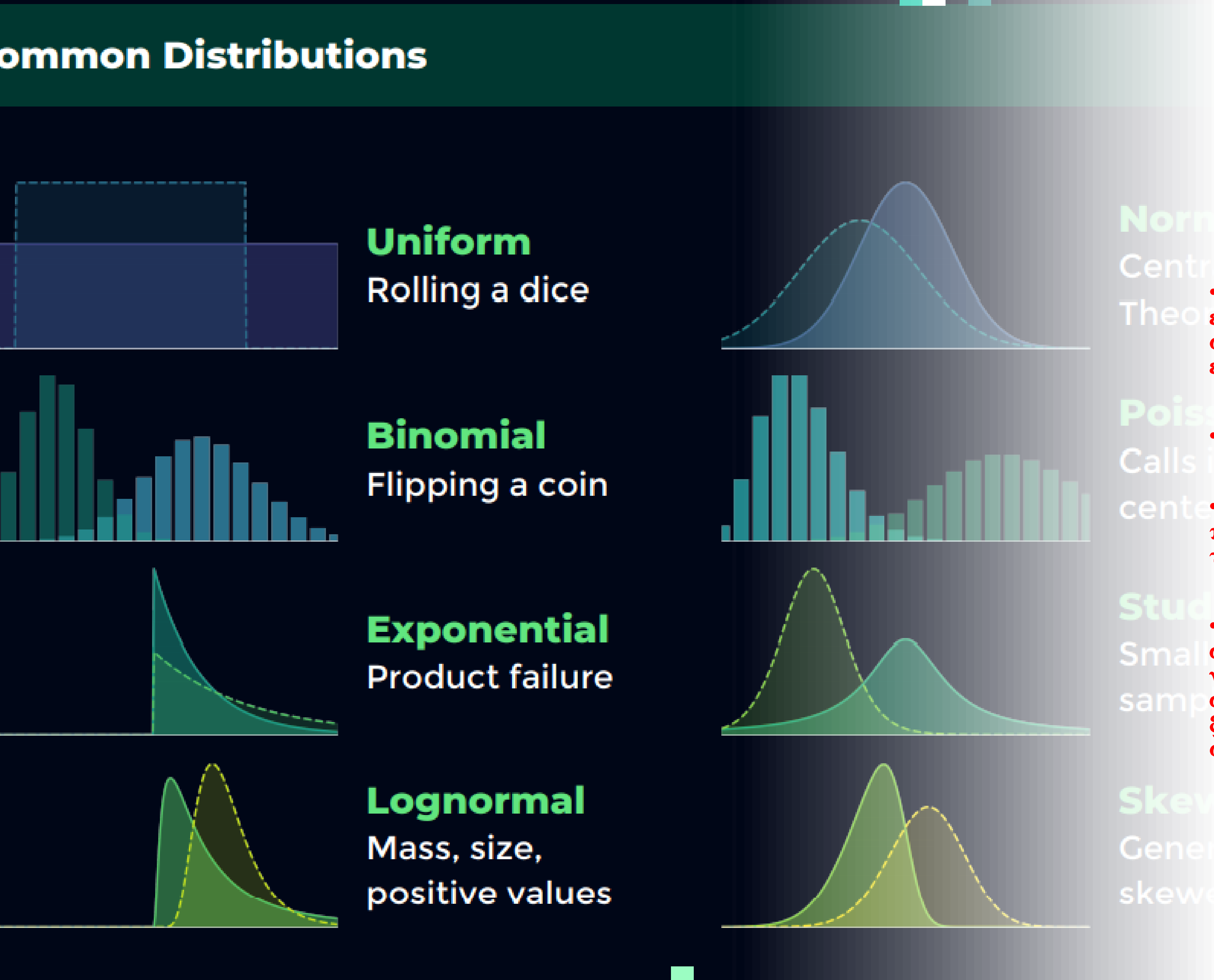
Student-t
Small-data samples



Lognormal
Mass, size, positive values



Skewnormal
Generic skewed dist



- Μπορούμε να έχουμε μια αρχική εκτίμηση της κατανομής από το σχήμα της αν δεν έχουμε δεδομένα για τον τρόπο που εκτελέσθηκε το πείραμα

- Δηλαδή

- Από μια άγνωστη χημική διαδικασία ποια κατανομή τα δεδομένα ακολουθούν τα δεδομένα

- Η αρχική έμπνευση υπάρχει από το σχήμα της κατανομής- φυσικά χρειάζεται να επιβεβαιωθεί η εκτίμησή μας με στατισιτκές δοκιμασίες – αλλά βοηθά αν ξέρουμε από που να βρούμε την πρώτη άκρη

Ιδιότητες μιας κατανομής (πιθανοτήτων)

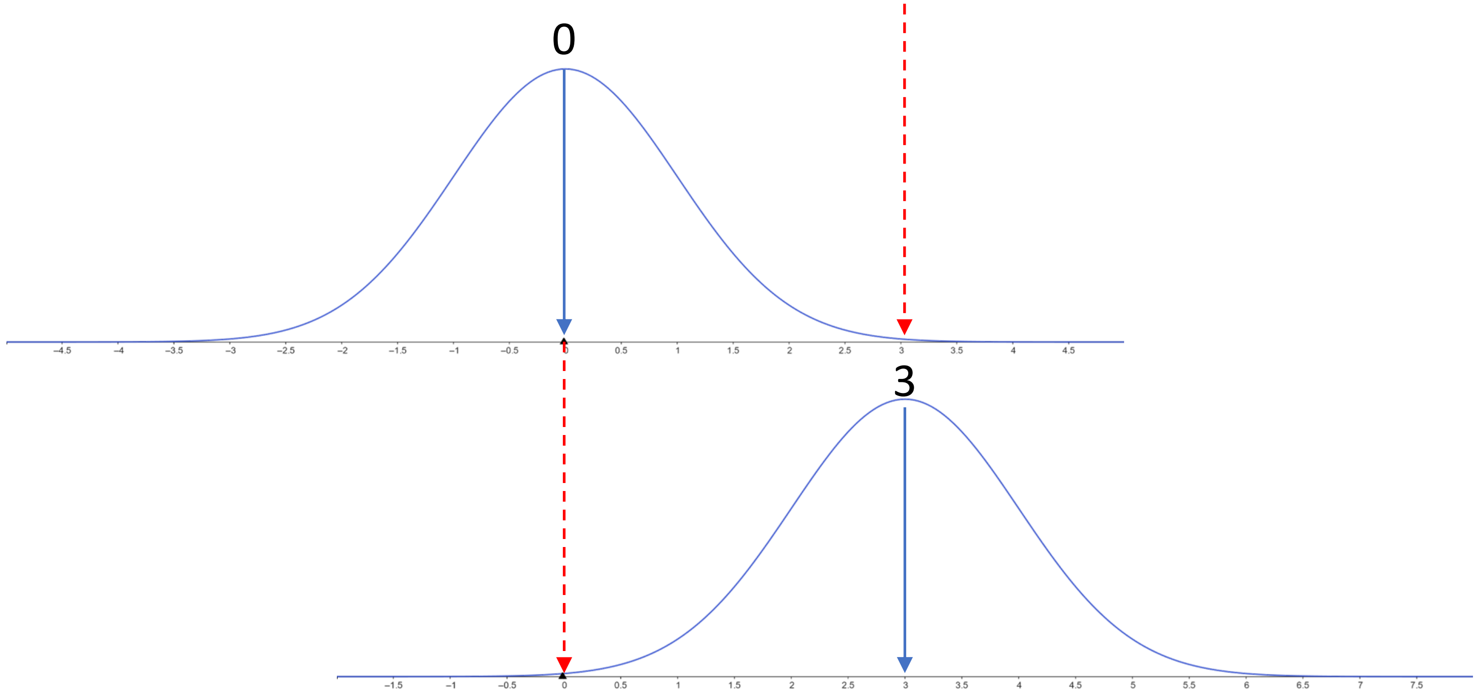
Επανάληψη

Δυο βασικές ιδιότητες χαρακτηρίζουν τις κατανομές

Μέση τιμή και διακύμανση

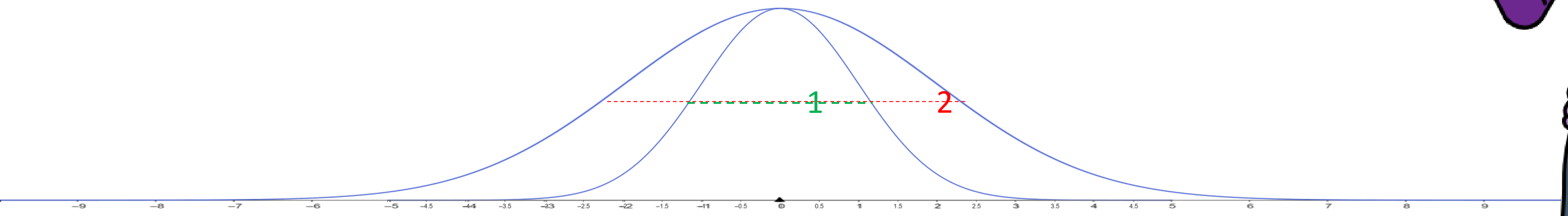
Ιδιότητες μιας κατανομής (πιθανότητων)

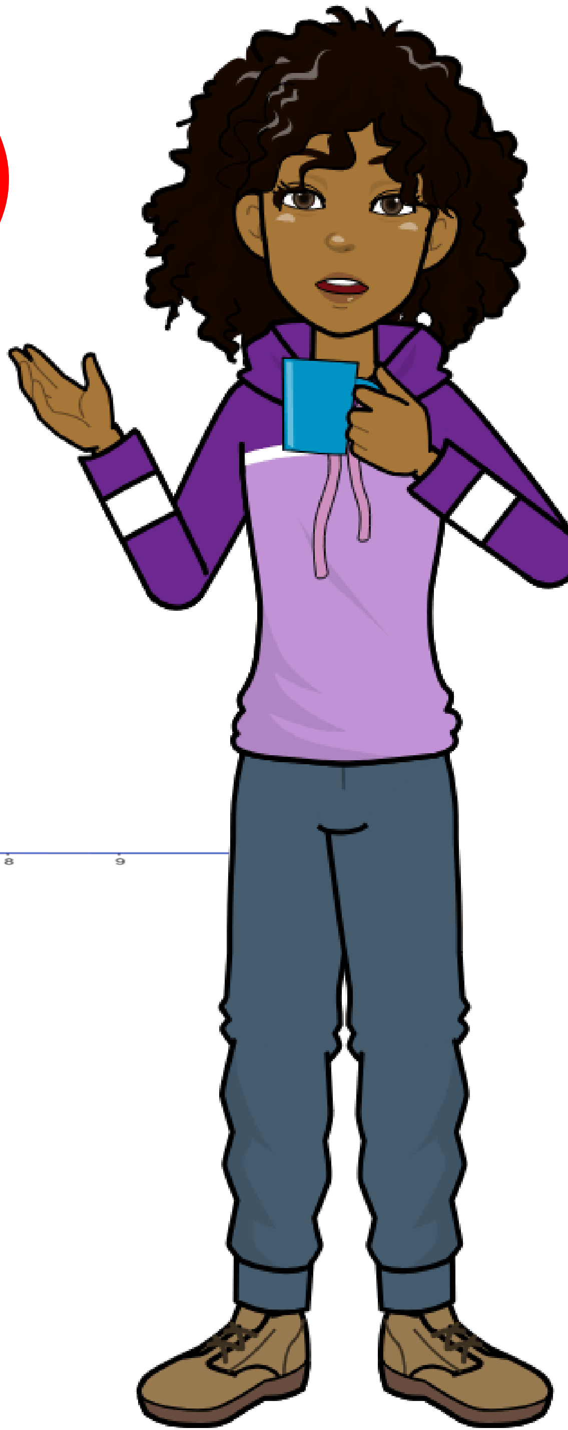
Μέση τιμή → στα μαθηματικά ο αριθμητικός μέσος όρος – είναι μέτρο θέσης – μας λέει που βρίσκεται η κατανομή μας – ποια τιμή έχει



Ιδιότητες μιας κατανομής (πιθανότητων)

Διακύμανση → πόσο στενά είναι διατεταγμένες οι τιμές γύρω από το μέσο όρο – μέτρο «πλάτους» που το λέμε μέτρο διασποράς





Ιδιότητες μιας κατανομής (πιθανότητων)

Μια διευκρίνηση

Αντιμετωπίζουμε διαφορετικά ένα πληθυσμό και διαφορετικά ένα δείγμα

Ο όρος πληθυσμός αναφέρεται στο σύνολο των αντικειμένων (των δεδομένων)

Ο όρος δείγμα αναφέρεται σε ένα μέρος των αντικειμένων → μέσω του δείγματος προσπαθούμε να προσεγγίζουμε τις ιδιότητες του πληθυσμού

Π.χ. μια παρτίδα φαρμάκων αποτελείται από 100.000 δισκία – Αυτός είναι ο πληθυσμός 100.000 δισκία

Το δείγμα είναι ένα μέρος των δισκίων που θα αναλύσουμε από τον πληθυσμό π.χ. 30 δισκία – ελπίζουμε (και προσπαθούμε να το διασφαλίσουμε με τη στατιστική) ότι το δείγμα είναι αντιπροσωπευτικό του πληθυσμού - δεν διαφέρουν στις ιδιότητες τους – π.χ. έχουν την ίδια κατανομή βάρους

Όμως τα συμβολίζουμε διαφορετικά άλλα σύμβολα ιδιοτήτων έχει ο πληθυσμός και άλλα το δείγμα πληθυσμός

δείγμα

Ιδιότητες μιας κατανομής (πιθανότητων)

Μέση τιμή

Η μέση τιμή του πληθυσμού συμβολίζεται ως μ - Η μέση τιμή του δείγματος ως $\bar{x}$

Διακύμανση

Η διακύμανση του πληθυσμού συμβολίζεται ως σ^2 - Η διακύμανση του δείγματος ως s^2

Ιδιότητες μιας κατανομής (πιθανοτήτων)

Η διακύμανση μετριέται σε δείγμα^2 – οδηγεί σε περίεργα αποτελέσματα – πχ μετρήσεις σε g^2

Για το λόγο αυτό εισάγεται μια νέα εκτιμήτρια συνάρτηση – η τυπική απόκλιση → $\sqrt{\sigma^2} = \sigma$ που ονομάζεται (πληθυσμιακή) τυπική απόκλιση και $\sqrt{s^2} = s$ που ονομάζεται (δειγματική) τυπική απόκλιση

Συσχέτιση μεταξύ σ^2 και μ

$$\sigma^2 = \mathrm{E}\left((\Upsilon - \mu)^2\right)$$

Όπου Ε ονομάζεται αναμενόμενη τιμή (expectation value) και είναι ο ζυγισμένος με τις πιθανότητες μέσος όρος (ακριβέστερα η πιο πιθανή τιμή που θα πάρουμε ως αποτέλεσμα – αντιληπτικά δεν είναι ο μέσος όρος που προέρχεται από πολλές τιμές αλλά μια τιμή που προβλέπει το πιθανότερο αποτέλεσμα του επόμενου πειράματος – έχουν όμως την ίδια αριθμητική τιμή – ας πούμε ότι η αναμενόμενη τιμή είναι a priori πρόβλεψη και ο μέσος όρος a posteriori γνώση μετά την εκτέλεση πειραμάτων)

$$E(x) = \sum x\, P(x)$$

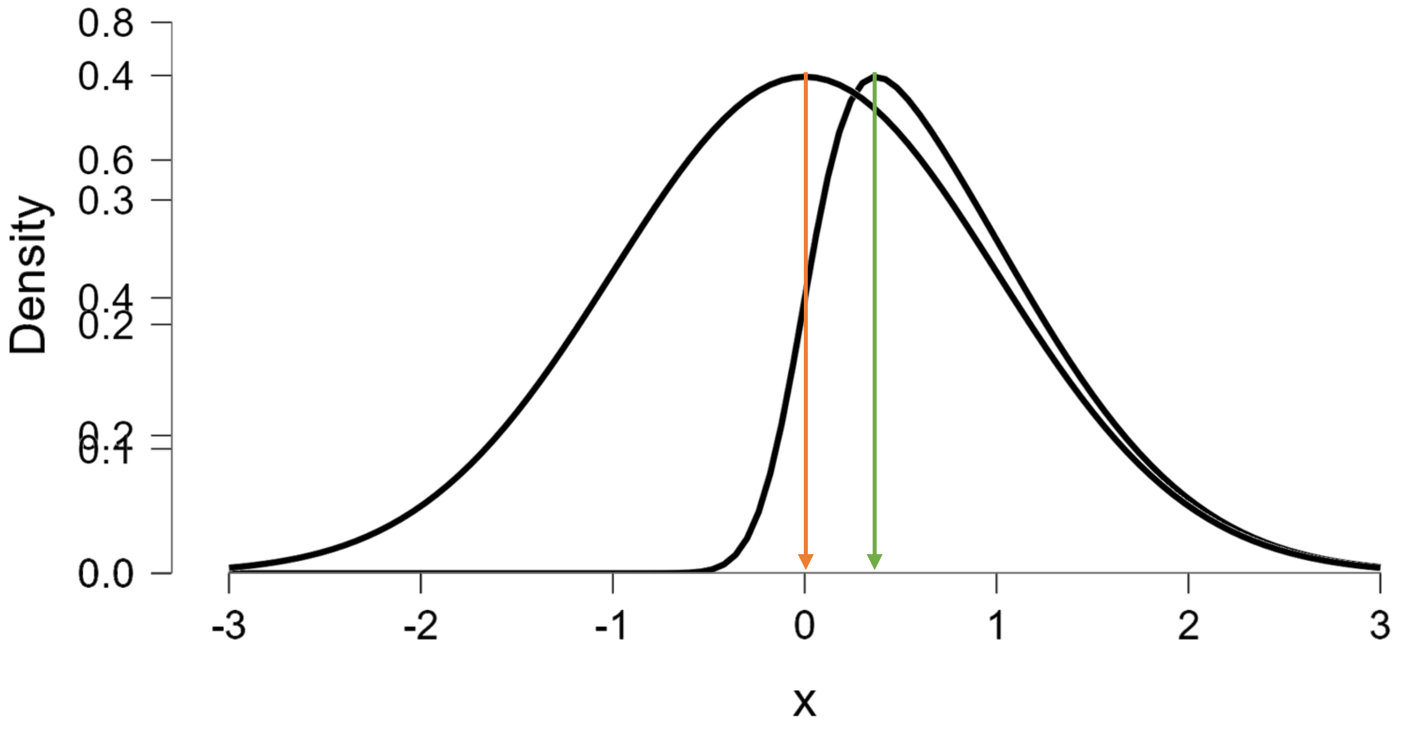
Ιδιότητες μιας κατανομής (πιθανοτήτων)

Στην αρχή είχαμε πει τρεις ιδιότητες… Είναι αλήθεια? Υπάρχει $3^{η}$ ιδιότητα?

Ιδιότητες μιας κατανομής (πιθανοτήτων)

Είναι αλήθεια Υπάρχει $3^{η}$ ιδιότητα. Η κύρτωση που φαίνεται από τη συμμετρία

Ιδιότητες μιας κατανομής (πιθανοτήτων)



Η λοξότητα (skewness) είναι το πόσο συμμετρική είναι η κατανομή.
Συνήθως εκτιμάται στη μέση του ύψους

$$Pearson\ skewness = 3 \times \frac{\mu - \delta\iota\alpha\mu\varepsilon\sigma\eta\ \tau\iota\mu\eta}{\sigma}$$

Θα δούμε διάφορες κατανομές

Σε κάθε κατανομή υπάρχουν δύο ειδών διαγράμματα

Η συνάρτηση πυκνότητας πιθανότητας (probability density function - pdf) και η αθροιστική συνάρτηση πυκνότητας πιθανότητας (cumulative density function - cdf)

συνάρτηση πυκνότητας πιθανότητας ο συνολικός αριθμός πιθανοτικών περιστατικών για κάθε τιμή

ομοιόμορφη κατανομή (uniform distribution)

Κατανομή όπου κάθε αποτέλεσμα έχει την ίδια πιθανότητα να παρατηρηθεί

$$f(x) = \begin{cases} \frac{1}{b-a} & , a \leq x \leq b, \\ 0 & , x < a \text{ or } x > b \end{cases}$$

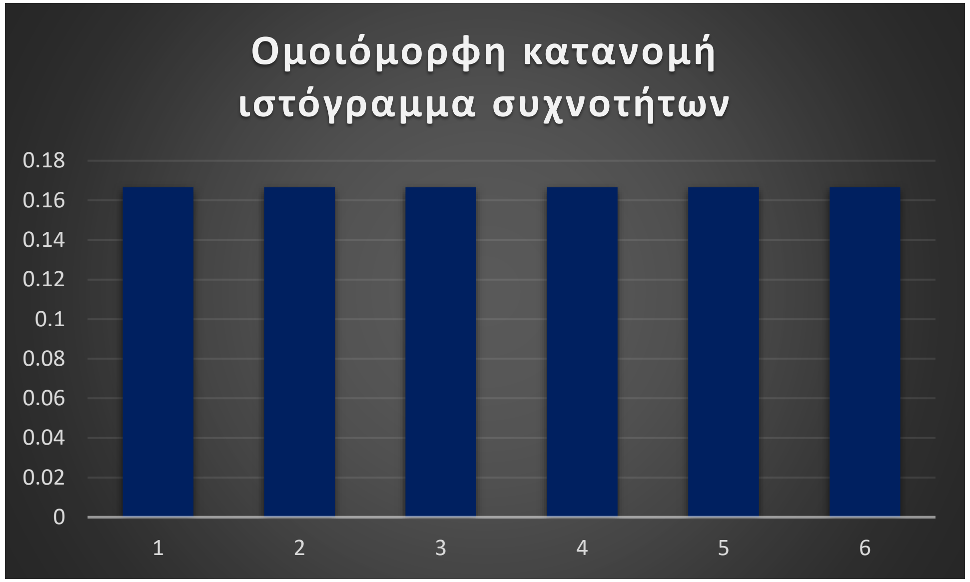
Αν το χ βρίσκεται μεταξύ δυο ορίων a και b η πιθανότητα να παρατηρηθεί ένα αποτέλεσμα είναι 1/b-a

Αν το χ βρίσκεται εκτός δυο ορίων a και b η πιθανότητα είναι 0

Η ολική πιθανότητα είναι 1 και το πλάτος είναι b-a

1=(b-a) x p(a) <=> p(a)=1/b-a

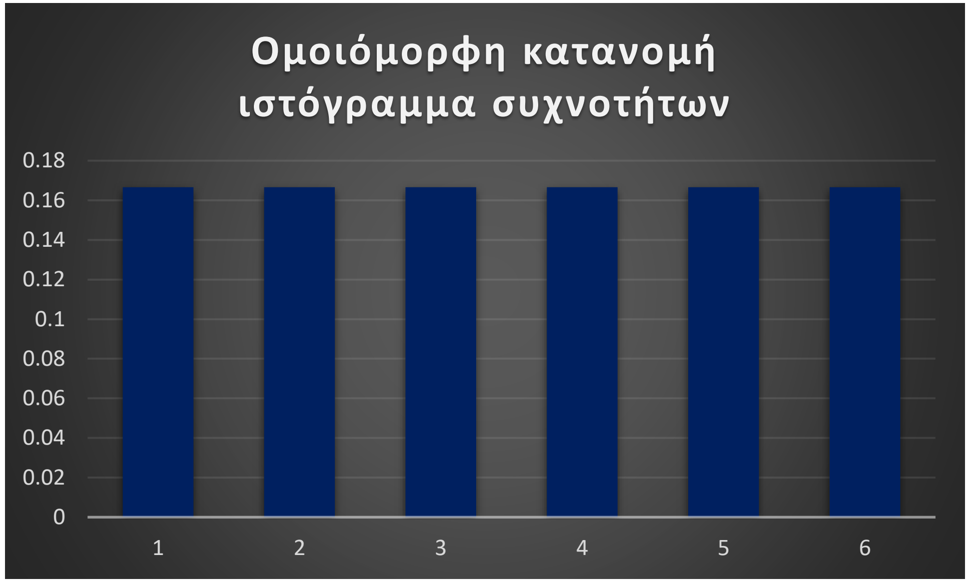
Το Γράφημα της ομοιόμορφης κατανομής είναι ένα άθροισμα παραλληλόγραμμων



Ο μέσος όρος είναι $\mu = \frac{a+b}{2}$

Η διασπορά είναι $\sigma = \frac{(b-a)^2}{12}$

Το Γράφημα της κανονικής κατανομής είναι ένα άθροισμα παραλληλόγραμμων



Ο μέσος όρος είναι $\mu = \frac{a+b}{2}$

Η διασπορά είναι $\sigma = \frac{(b-a)^2}{12}$

Η ομοιόμορφη κατανομή είναι γενικά μικρής χρησιμότητας (ως και πολύ μικρής)

Δεν δίνει πολλές πληροφορίες πέραν εκείνων που διαισθητικά διαθέτουμε

ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Ονομάζεται διωνυμική γιατί μπορεί να πάρει μόνο δύο τιμές όπως πχ στο στρίψιμο ενός νομίσματος,κλπ

Όταν κάνουμε το πείραμα μία και μόνη φορά
p πιθανότητα να συμβεί ένα φαινόμενο (επιτυχία) άρα η πιθανότητα να μην συμβεί είναι 1-P=q (αποτυχία)
Αν επαναλάβουμε το πείραμα n φορές, η πιθανότητα P να συμβεί το φαινόμενο x φορές (άρα να μην συμβεί n-x)
δίνεται από τη σχέση

$$P(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

Όταν κάνουμε το πείραμα μια φορά η κατανομή πιθανοτήτων ονομάζεται **Bernouli** (δεν θα ασχοληθούμε)

Πως αναγνωρίζουμε μια διωνυμική κατανομή?

1. Δύο ΜΟΝΟ αποτελέσματα (σωστό-λάθος, 0-1, pass-fail κλπ)
2. Ανεξαρτησία πειραμάτων (αν το ν-1 πείραμα έχει ως αποτέλεσμα λάθος, το ν θα διατηρήσει την ελευθερία του να βγάλει σωστό / λάθος, δεν θα επηρεαστεί από το προηγούμενο πείραμα)
3. Καθορισμένος (μικρός) αριθμός επαναλήψεων n
4. Η πιθανότητα επιτυχίας p ενός αποτελέσματος παραμένει σταθερή

1° ΚΕΦΑΛΑΙΟ

1. Το μικρότερο σωματίδιο της ύλης που μπορεί να υπάρχει σε ελεύθερη κατάσταση και να διατηρεί τις ιδιότητες της ουσίας στην οποία ανήκει είναι :
α. το άτομο
β. το μόριο
γ. το ιόν
δ. το πρωτόνιο

2. Τα ιόντα είναι:
α. ηλεκτρικά φορτισμένα σωματίδια
β. ηλεκτρικά φορτισμένα άτομα
γ. ηλεκτρικά φορτισμένα συγκροτήματα ατόμων
δ. άτομα ή συγκροτήματα ατόμων με ηλεκτρικό φορτίο.

κατανομή Bernoulli

3. Ο ατομικός αριθμός εκφράζει:
α. τον αριθμό των πρωτονίων και νετρονίων στον πυρήνα ενός ατόμου.
β. τον αριθμό των ηλεκτρονίων ενός μονοατομικού ιόντος
γ. τον αριθμό των νετρονίων στον πυρήνα ενός ατόμου
δ. τον αριθμό των πρωτονίων στον πυρήνα κάθε ατόμου ενός στοιχείου

4. Ο μαζικός αριθμός εκφράζει:
α. τον αριθμό των πρωτονίων και νετρονίων στον πυρήνα ενός ατόμου.
β. τον αριθμό των ηλεκτρονίων ενός μονοατομικού ιόντος
γ. τον αριθμό των νετρονίων στον πυρήνα ενός ατόμου
δ. τον αριθμό των πρωτονίων στον πυρήνα κάθε ατόμου ενός στοιχείου

5. Το κατιόν Ca^{2+} περιέχει 20 νετρόνια και 18 ηλεκτρόνια. Ο μαζικός αριθμός του Ca είναι :
α. 40 β. 38 γ. 20 δ. 18

6. Το ανιόν Cl^- περιέχει 20 νετρόνια και 18 ηλεκτρόνια. Ο μαζικός αριθμός του Cl είναι :
α. 20 β. 37 γ. 38 δ. 17

7. Τα ισότοπα άτομα έχουν:
α. ίδιο αριθμό πρωτονίων και νετρονίων
β. ίδιο μαζικό και διαφορετικό ατομικό αριθμό
γ. ίδιο αριθμό πρωτονίων και διαφορετικό αριθμό νετρονίων
δ. ίδιο αριθμό πρωτονίων και διαφορετικό αριθμό ηλεκτρονίων.

8. Σε ποια από τις παρακάτω περιπτώσεις δε θα σχηματιστεί μείγμα;
α. κατά την προσθήκη ζάχαρης σε νερό
β. κατά την προσθήκη νερού σε λάδι
γ. κατά την ανάμειξη ζεστού με κρύο νερό
δ. κατά το επιφανειακό σκούριασμα του σιδήρου

9. Ποια από τις παρακάτω ιδιότητες που αναφέρονται στα ομογενή μείγματα δεν ισχύει ;
α. έχουν ίδια πυκνότητα σε όλη την έκταση του όγκου τους
β. έχουν μεταβλητή πυκνότητα, ανάλογα με την αναλογία με την οποία αναμείχτηκαν τα συστατικά τους
γ. η πυκνότητα τους ισούται με το άθροισμα των πυκνοτήτων των συστατικών τους
δ. η πυκνότητά τους αυξάνεται όταν ψύχονται με σταθερή πίεση.

διωνυμική κατανομή

Ένα τεστ ερωτήσεων πολλαπλής επιλογής

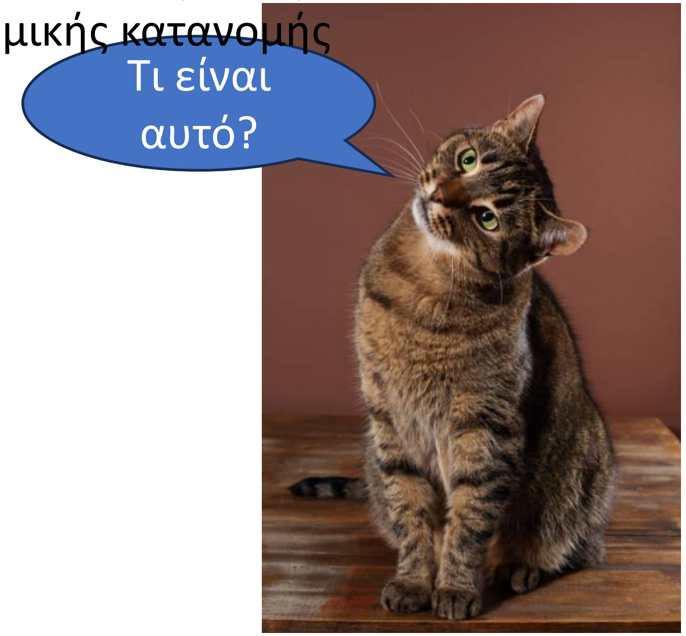
- Σε κάθε μια ερώτηση ακολουθεί (ποια είναι η πιθανότητα να απαντήσουμε σωστά σε μία από τις ερωτήσεις)
- Στο σύνολο ακολουθεί (ποια είναι η πιθανότητα να απαντήσουμε σωστά στο 70% των ερωτήσεων του τεστ)

Ονομάζεται διωνυμική γιατί μπορεί να πάρει μόνο δύο τιμές όπως πχ στο στρίψιμο ενός νομίσματοσκλπ

Όταν κάνουμε το πείραμα μία και μόνη φορά
p πιθανότητα να συμβεί ένα φαινόμενο (επιτυχία) άρα η πιθανότητα να μην συμβεί είναι 1-P=q (αποτυχία)
Αν επαναλάβουμε το πείραμα n φορές, η πιθανότητα P να συμβεί το φαινόμενο x φορές (άρα να μην συμβεί n-x) δίνεται από τη σχέση

$$P(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

Που είανι η συνάρτηση πιθανότητας της διωνυμικής κατανομής



Ας θέσουμε το ερώτημα

Ένα κέρμα
2

1
Ποια είναι η πιθανότητα

3
Μετά από n επαναλήψεις

$$P(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

6
ξέροντας ότι σε μια επανάληψη έρχεται
p φορές από τη μεριά του κεφαλιού

και q φορές από την άλλη μεριά
7

να έλθει x φορές από τη μία πλευρά

4

(άρα n-x φορές από την άλλη)

5

Ας θέσουμε το ερώτημα

Ο όρος αυτός μας δείχνει τους συνδυασμούς που μπορεί να υπάρχουν

ξέροντας ότι σε μια επανάληψη έρχεται p φορές από τη μεριά του κεφαλιού

$$P(x) = \frac{n!}{x!\,(n-x)!}p^x q^{n-x}$$

και q φορές από την άλλη μεριά

π.χ. Τρεις συνδυασμοί για το παραπάνω κέρμα με ένα κεφάλι και δύο γράμματα

$$\frac{n!}{x!(n-x)!} = \frac{3!}{1!(3-1)!} = \frac{3!}{1!(3-1)!}$$

3 συνδυασμοί

Ας θέσουμε το ερώτημα

Ο όρος αυτός μας δείχνει τους συνδυασμούς που μπορεί να υπάρχουν

ξέροντας ότι σε μια επανάληψη έρχεται p φορές από τη μεριά του κεφαλιού

$$P(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

και q φορές από την άλλη μεριά

Αν χρησιμοποιήσω ένα κανονικό κέρμα οι πιθανότητες είναι 0.5 για κάθε πλευρά
Θα χρησιμοποιήσω ένα «πειραγμένο» που οι πιθανότητες να έλθει κεφάλι είναι 40%
Άρα p=0.4 συνεπώς q=1-p=0.6
Αν κάνω 4 προσπάθειες ποια είναι η πιθανότητα να φέρω 2 φορές (από τις 4) κεφάλι?

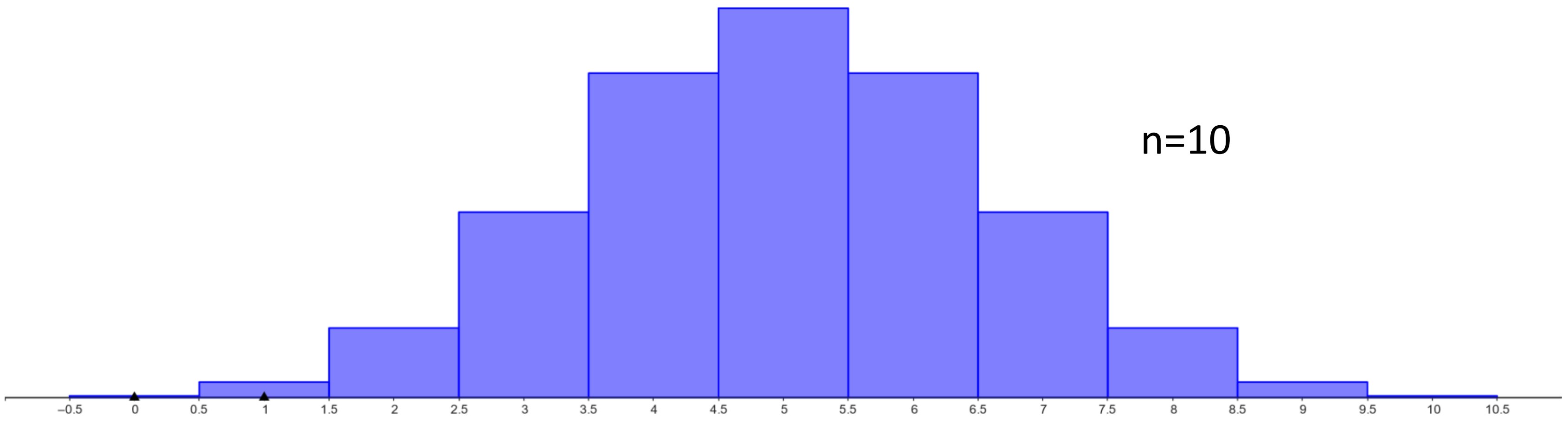
$$p^x q^{n-x} = 0.4^2 0.6^{4-2} = 0.4^2 0.6^{4-2} = 0.4 \text{ x } 0.4 \text{ x } 0.6 \text{ x } 0.6$$

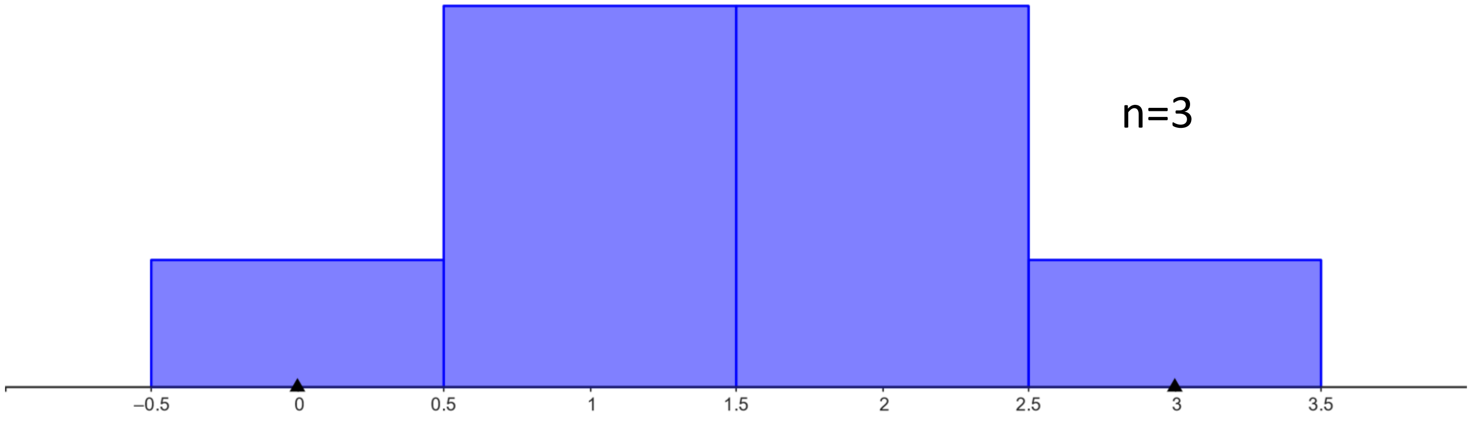
Ίσως πρέπει να βάλω 4 πειράματα και στην προηγούμενη διαφάνεια

Τι λεω? Ότι η πιθανότητα είναι το 0.4 του 0.4 του 0.6 του 0.6
Το δεύτερο μέρος μας λέει τις πιθανότητες και το πρώτο? Τους συνδυασμούς των πιθανοτήτων π.χ. μα άλλη πιθανότητα είναι το 0.4 του 0.6 του 0.4 του 0.6

Διάγραμμα της διωνυμικής κατανομής

Συνήθως χρησιμοποιούμε ένα ιστόγραμμα συχνοτήτων (είναι μη συνεχής=διακριτή κατανομή)
Αν πραγματοποιήσουμε n πειράματα θα παρασταθούν από n+1 ράβδους





Διακύμανση διωνυμικής κατανομής

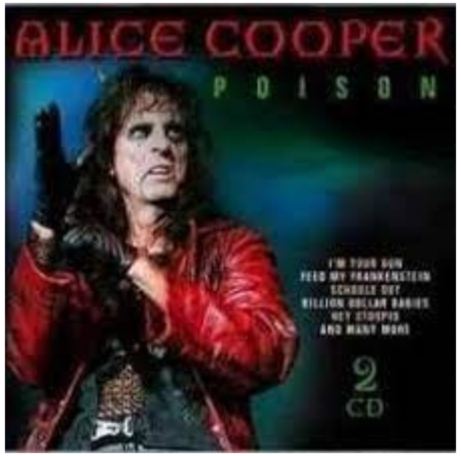
$\sigma^2 = npq = np(1-p)$

Μέσος όρος της διωνυμικής κατανομής

$\mu = np$

KATANOMH Poisson

KATANOMH Poisson

OXI Poison  ή 

pwa·son

https://www.youtube.com/watch?v=HeyYyfPgMmg&ab_channel=JulienMiquel

Η κατανομή Poisson περιγράφει την πιθανοτητα εμφάνισης ενός αποτελέσματος σε μια συγκεκριμένη ΣΥΝΕΧΗ ποσότητα (όγκος, χρόνος, απόσταση κλπ) αν ξέρουμε τη συχνότητα εμφάνισης του φαινομένου

Πχ σεισμοί στη διάρκεια ενός έτους, πιθανότητα εμφάνισης ενός συγκεκριμένου μορίου σε μια επιφάνεια που σαρώνει ένα φασματόμετρο μάζας MALDI-TOF κλπ

Η πιθανότητας εμφάνισης αυτού του αποτελέσματος έχει κάποιο συγκεκριμένο ρυθμό που έχουμε παρατηρήσει και συμβολίζεται με

$$\lambda$$

Τι θέλουμε να μάθουμε? Ποια είναι η πιθανότητα αυτό το φαινόμενο να συμβεί περισσότερες ή λιγότερες φορές από την συχνότητα εμφάνισης λ.
Π.χ. ποια η πιθανότητα εμφάνισης τριών σεισμών την επόμενη χρονιά, όταν ο ρυθμός εμφάνισής τους είναι 2/έτος

Προυποθέσεις για να υπάρχει κατανομή Poisson

1. Τα γεγονότα να είναι ανεξάρτητα μεταξύ τους πχ. στο πολύ ευχάριστο θέμα των σεισμών ο ένας να μην έχει προκαλέσει τον επόμενο
2. Ο ρυθμός εμφάνισης του φαινομένου λ να είναι σταθερός

Η συνάρτηση πιθανότητας (κανονικά για όλες τις κατανομές συνάρτηση πυκνότητας πιθανότητας Probability density function PDF) δίνεται από τον ακόλουθο τύπο

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Όπου x=0, 1, 2, 3, 4, 5…

Η κατανομή έχει μια ελάχιστη τιμή, το 0, αλλά στις θετικές τιμές →∞ συνεπώς ο συντελεστής ασυμμετρίας (**skewness**) είναι >1

Τι είναι το λ είπαμε?

Ο ρυθμός εμφάνισης ενός αποτελέσματος πχ πόσα αυτοκίνητα περνούν από ένα δρόμο σε μια ώρα

$$\lambda = \frac{αριθμός\ αυτοκινητων}{ώρα}$$

Ποιο είναι ο μέσος όρος (ή διασταλτικά λίγο η αναμενόμενη τιμή $E(x)$) της κατανομής Poisson?

$$\mu = \mathrm{E(x)} = \lambda$$

Ποια είναι η διασπορά σ^2 ?

$$\sigma^2 = \lambda$$

!!!!!!!!!!!!!!!!!!!

Η τυπική απόκλιση ?

$$\sigma = \sqrt{\sigma^2} = \sqrt{\lambda}$$

Σχέση της κατανομής Poisson με τη διωνυμική κατανομή

Η αναμενόμενη τιμή $E(x)$ της διωνυμικής κατανομής είναι $E(x) = np$

Εάν το n γίνει πολύ μεγάλο και το p γίνει πολύ μικρό τότε το $\lim \to \infty$ της διωνυμικής κατανομής είναι η κατανομή Poisson

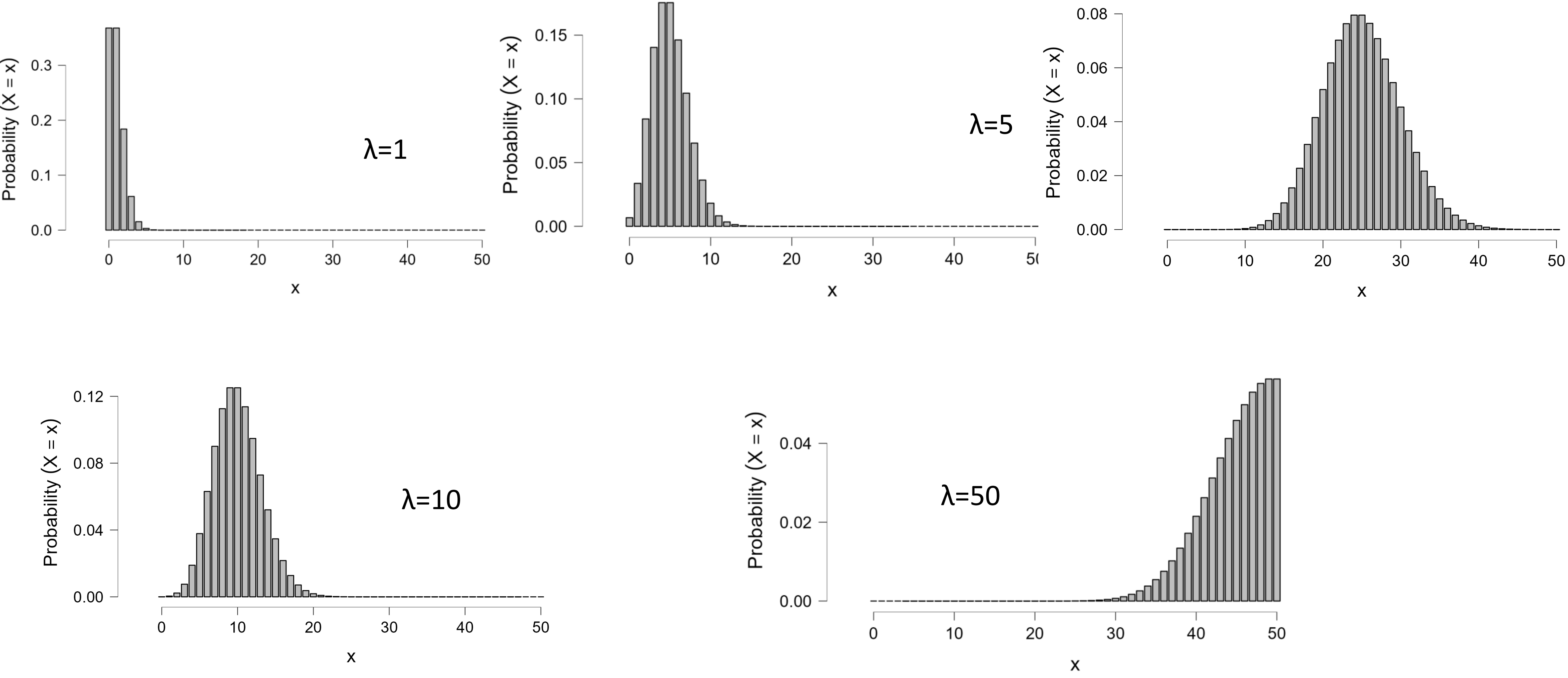
Ας δούμε ένα παράδειγμα

Όπως είπαμε πριν πχ πόσα αυτοκίνητα περνούν από ένα δρόμο σε μια ώρα και ας το προσεγγίσουμε με μια διωνυμική κατανομή $E(x) = np$ όπου

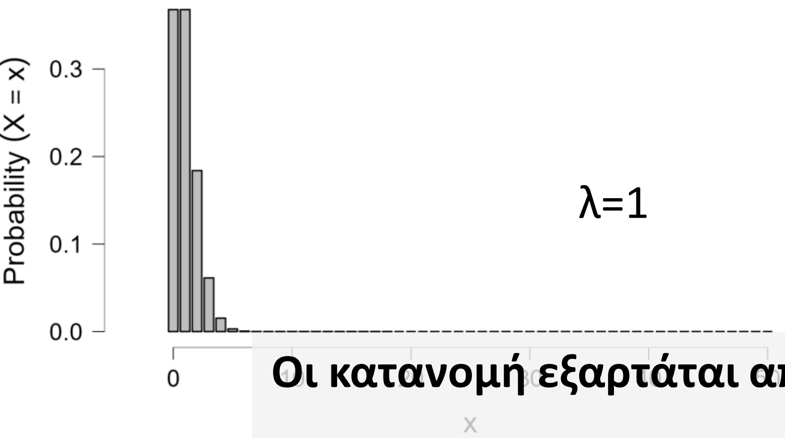
$n = \frac{αριθμός\ αυτοκινητων}{ώρα}$ και p=πιθανότητα να συμβαίνει το φαινόμενο αυτό

Αν ήθελα να έχω μια καλύτερη εκτίμηση ίσως θα ήταν πιο λογικό να μειώσω το χρόνο παρατήρησης πχ να τον κάνω 1 λεπτό.
Αυτό σημαίνει ότι το n θα μεγαλώνει και αντίστοιχα θα μειώνει την πιθανότητα
Ας πούμε ότι αρχικά ο αριθμός των παρατηρήσεων είναι $n = 10$ (δηλαδή έχω ξοδέψει 10 ώρες) και η πιθανότητα να παρατηρήσω 60 αυτοκίνητα είναι 0.5 (είπαμε είναι δυαδικό φαινόμενο η διωνυμική κατανομή)βλέπω η όχι 60 αυτοκίνητα
Αν αυξήσω τον αριθμό των παρατηρήσεων (κάθε λεπτό, άρα n=60x10=600) η πιθανότητα π μικραίνει – σιγά μην δω 60 αυτοκίνητα /λεπτό με πιθανότητα 0.5. Στη σχέση το γινόμενο όμως παραμένει περίπου σταθερό. Αυτή είναι και η ψυχή της κατανομής Poisson –είναι μια ειδική διωνυμική κατανομή

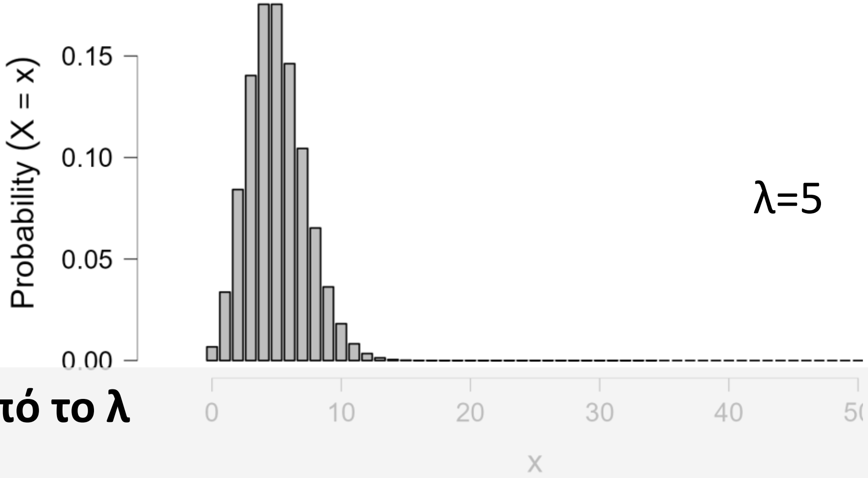
Πως μοιάζει ένα διάγραμμα κατανομής Poisson? Για n=50



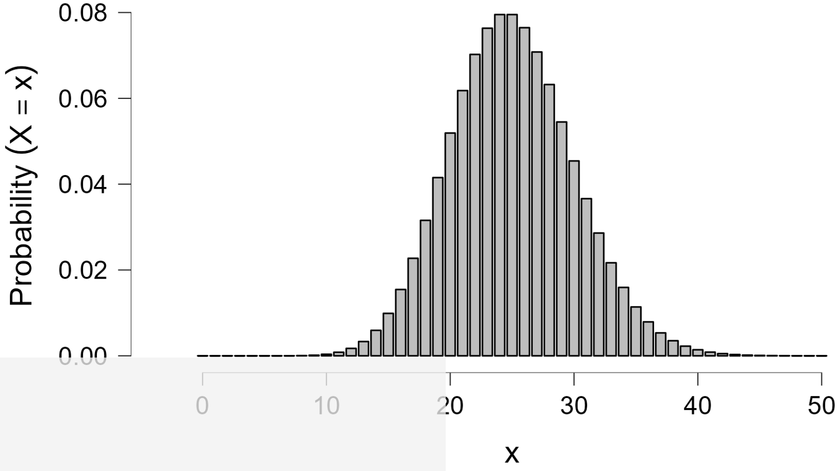
Πως μοιάζει ένα διάγρμμα κατανομής Poisson? Για n=50



λ=1



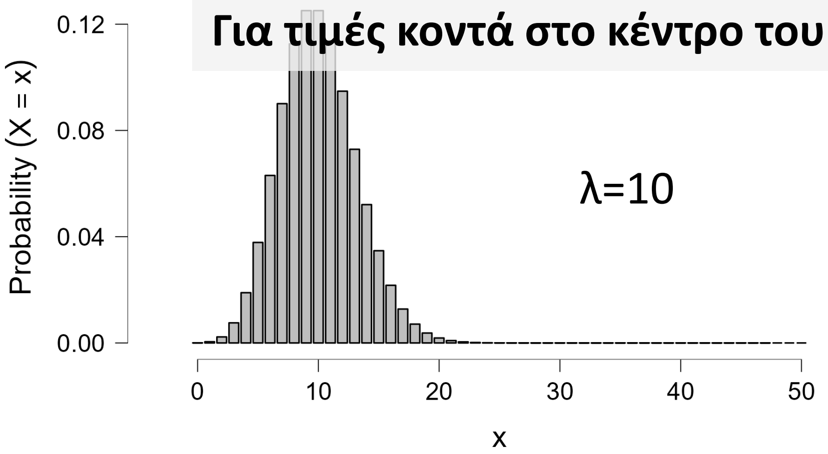
λ=5



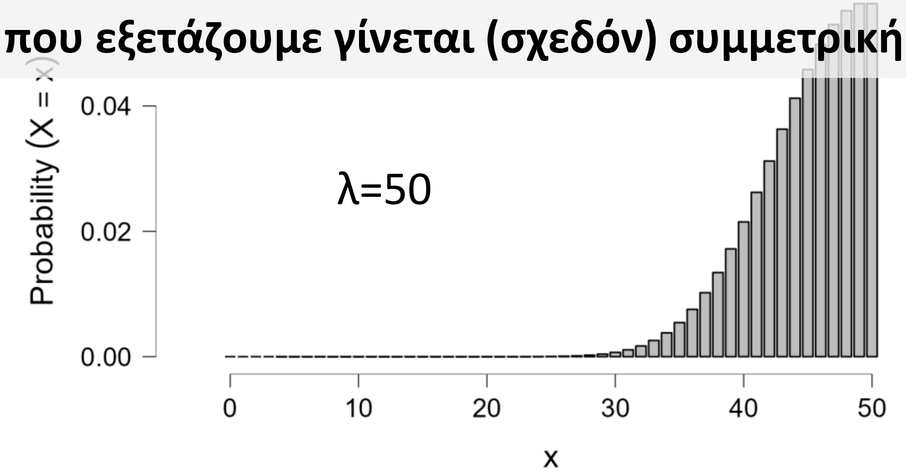
Οι κατανομή εξαρτάται από το λ

Η θέση της αλλά και το πλάτος της όπως έχουμε δει

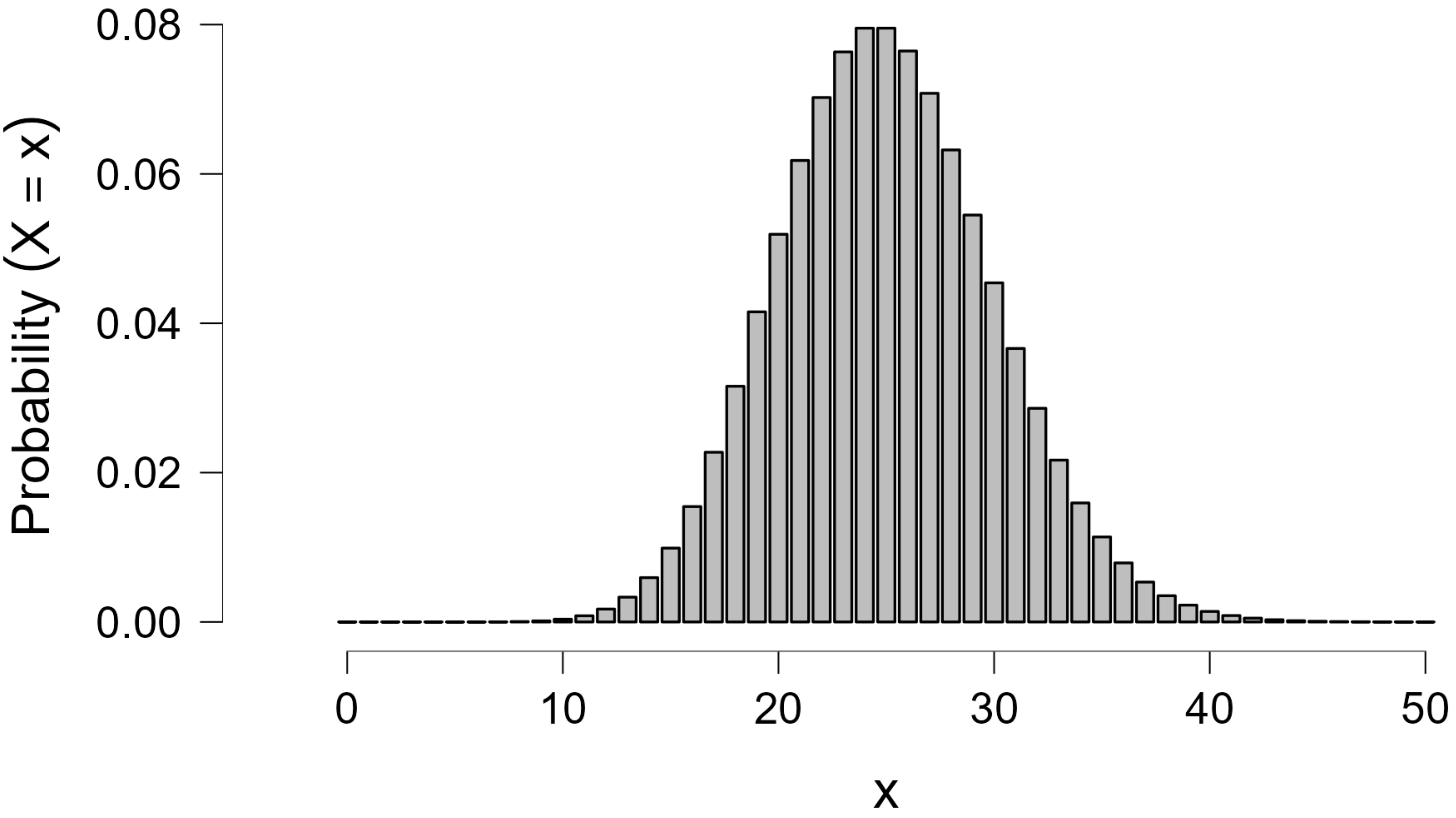
Για τιμές κοντά στο κέντρο του διαστήματος που εξετάζουμε γίνεται (σχεδόν) συμμετρική



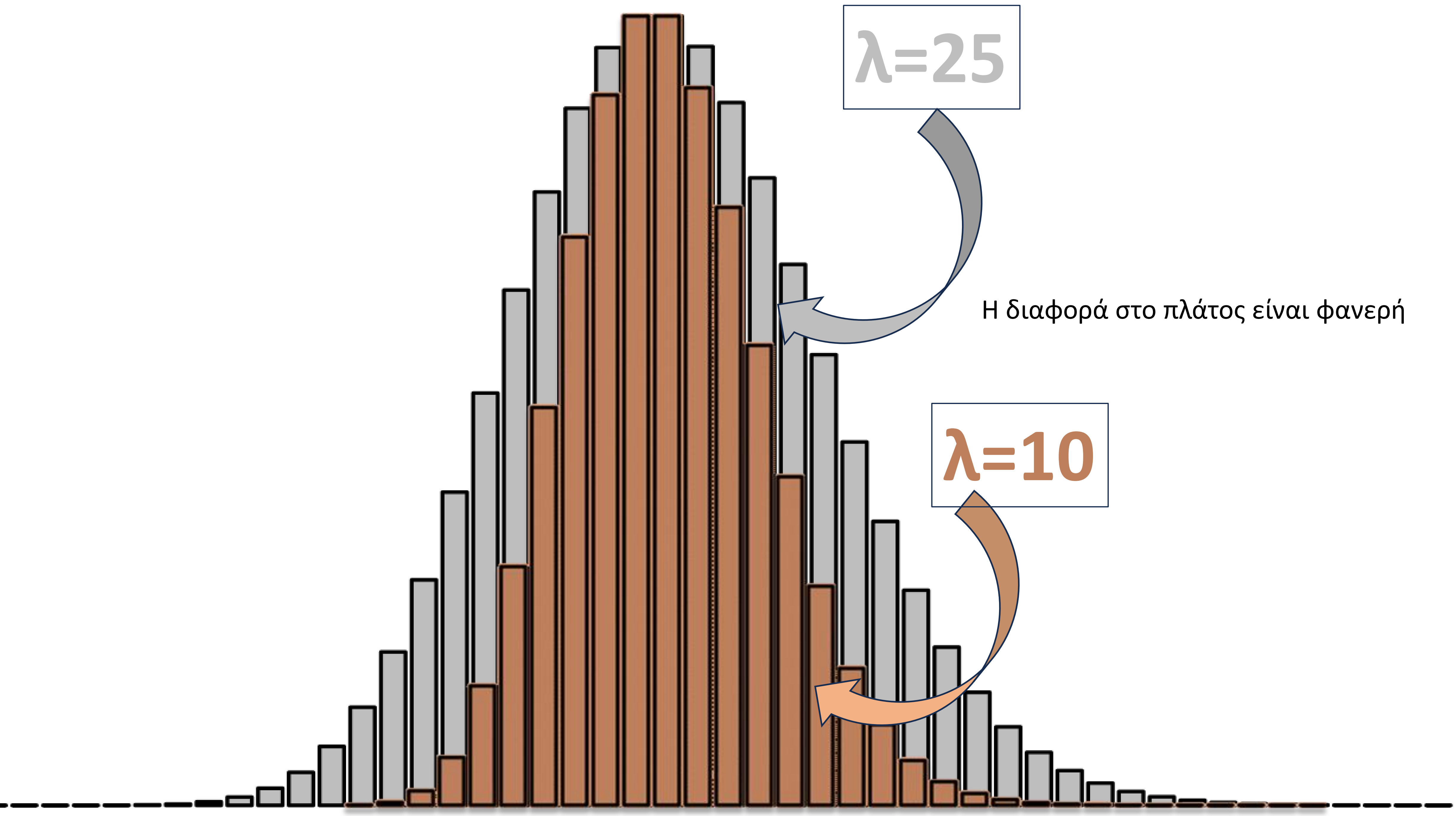
λ=10



λ=50



0.08
0.06
0.04
0.02
0.00
Probability (X = x)
0
10
20
30
40
50
x

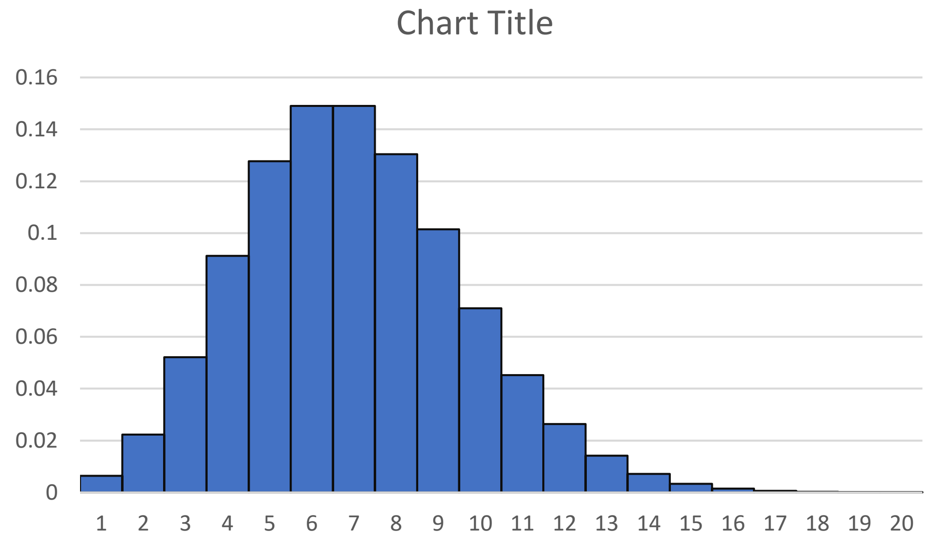

λ=25
Η διαφορά στο πλάτος είναι φανερή
λ=10

Ποια είναι η εντολή στο exel για να δημιουργήσω μια κατανομή?
Θα την δημιουργήσουμε μαζί και στο λογισμικό JASP

mean λ	7
P(x)	
1	0.006383
2	0.022341
3	0.052129
4	0.091226
5	0.127717
6	0.149003
7	0.149003
8	0.130377
9	0.101405
10	0.070983
11	0.045171
12	0.02635
13	0.014188
14	0.007094
15	0.003311
16	0.001448
17	0.000596
18	0.000232
19	8.54E-05
20	2.99E-05

=POISSON.DIST(A5,B2,FALSE)

P(x)   mean   Cumulative (FALSE, TRUE)



ΚΑΤΑΝΟΜΗ ΚΑΤΑ Gauss
Και
ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

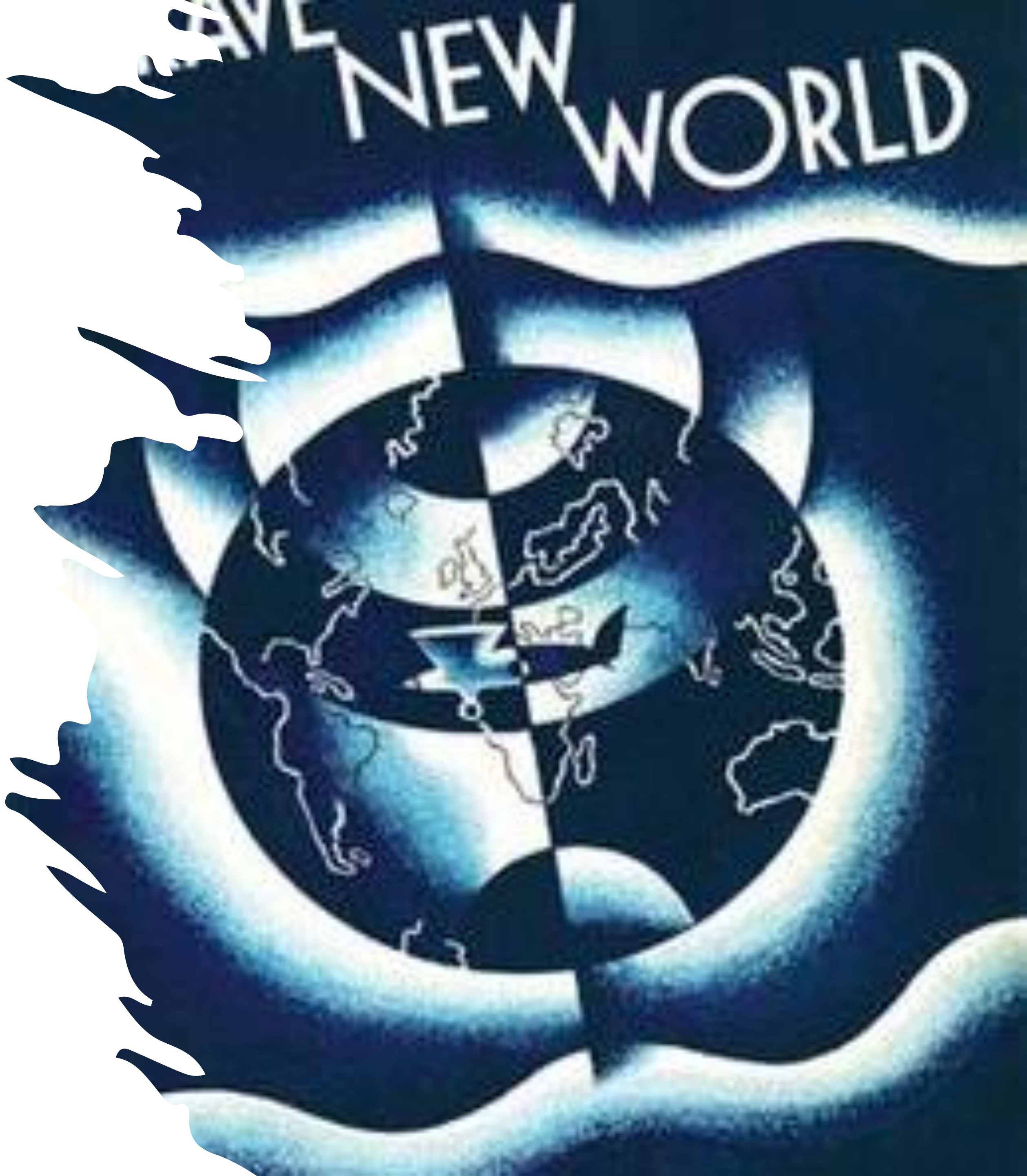
Η κατανομή κατά Gauss είναι η πιο χρησιμοποιούμενη κατανομή στις φυσικές επιστήμες

Ένας μεγάλος αριθμός στατιστικών δοκιμασιών υποθέτουν την ύπαρξη κατανομής Gauss στα δεδομένα και προσπαθούν να αποδείξουν ότι τα δεδομένα ακολουθούν κανονική κατανομή (normality tests)

Πολλές χημικές δοκιμασίες εμπλέκουν την κατανομή Gauss (πολύ σπανιότερα και σε δύσκολες περιπτώσεις την κατανομή Lorentz)

Π.χ.
Οι χρωματογραφικές κορυφές, οι κορυφές των φασμάτων μάζας και πολύ συχνά του NMR, τα σημεία σε μια καμπύλη αναφοράς κλπ προσεγγίζονται με κατανομές Gauss

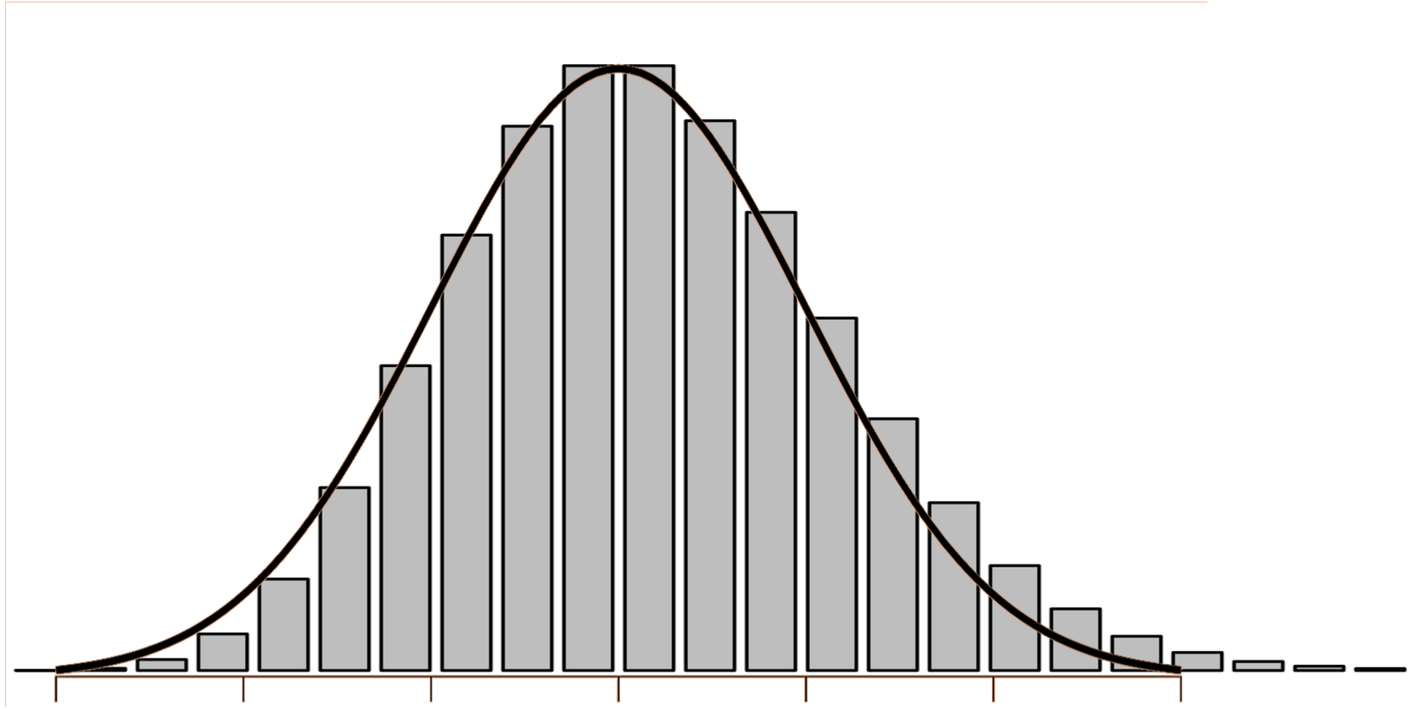
Η κατανομή Gauss είναι η πρώτη συνεχής (continoous) κατανομη που εξετάζουμε – δηλαδή μπορεί να πάρει απειρες τιμές (που προέρχονται από κάποια πολυπλοκη συχνά συνάρτηση)



Η ιδέα για την κατανομή Gauss προέρχεται από την κατανομή Poisson

Η γραφική παράσταση της κατανομής Poisson εξαρτάται από το λ

Για λ>10 η κατανομή Poisson γίνεται σχεδόν συμμετρική (δεν μπορεί να πάρει αρνητικές τιμές – ποτέ δεν γίνεται ΑΠΟΛΥΤΑ συμμετρική) και μοιάζει με την κατανομή Gauss που είναι πάντα συμμετρική



Πως αναγνωρίζουμε μια κατανομή Gauss? (στα ελληνικά συχνά βλέπουμε τον όρο Γκαουσιανή…)

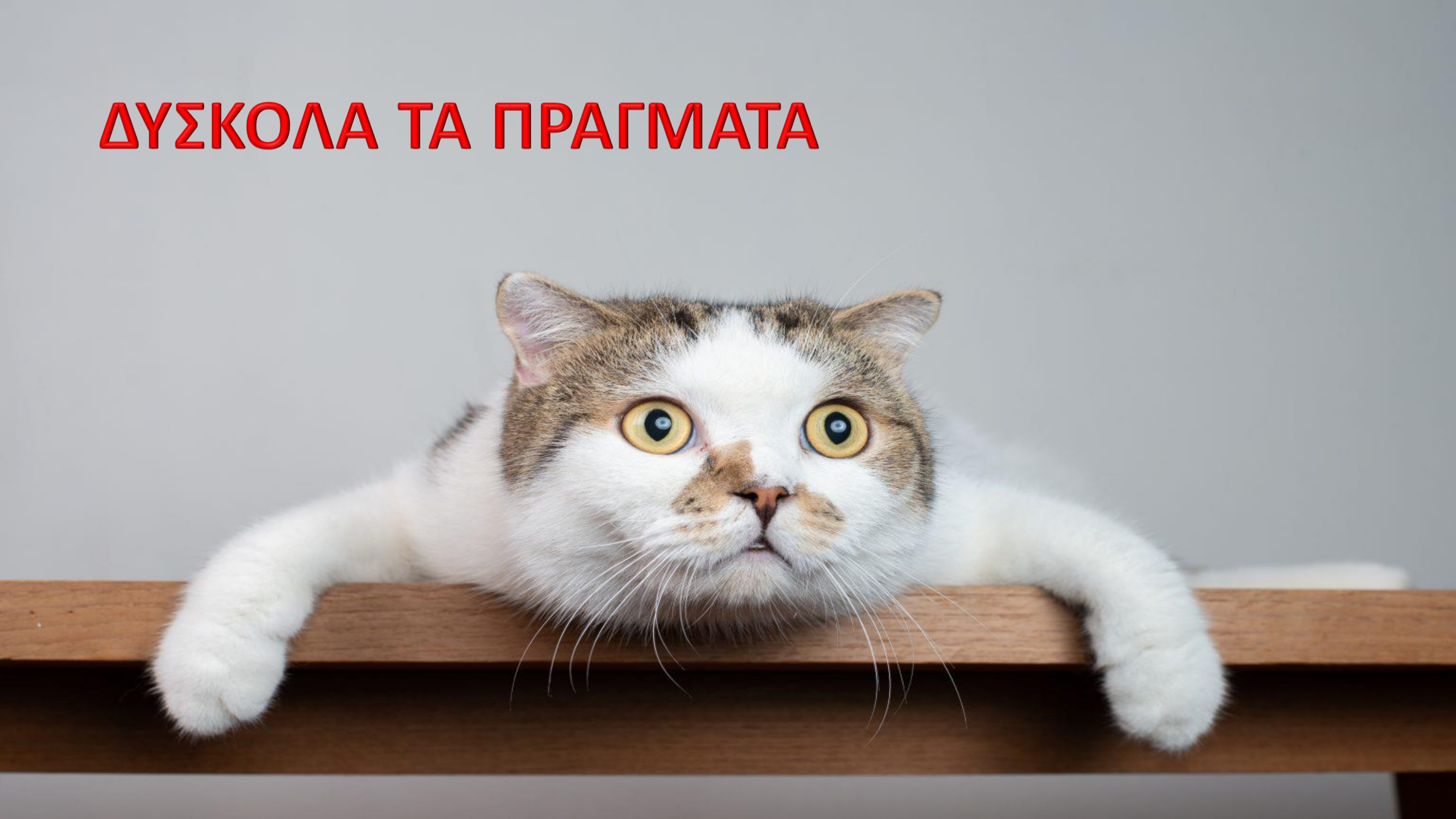
1. Είναι συνεχής συνάρτηση
2. Τα πειράματα είναι ανεξάρτητα
3. Έχει σχήμα καμπάνας
4. Έχει συμμετρία και έλλειψη κυρτότητας
5. Είναι ασυμπτωτική στον άξονα των x
6. Η μέση τιμή συμπίπτει με τη διάμεση
7. Ισχύει ο κανόνας 68-96-99.7

Κανονικότητα (normality)

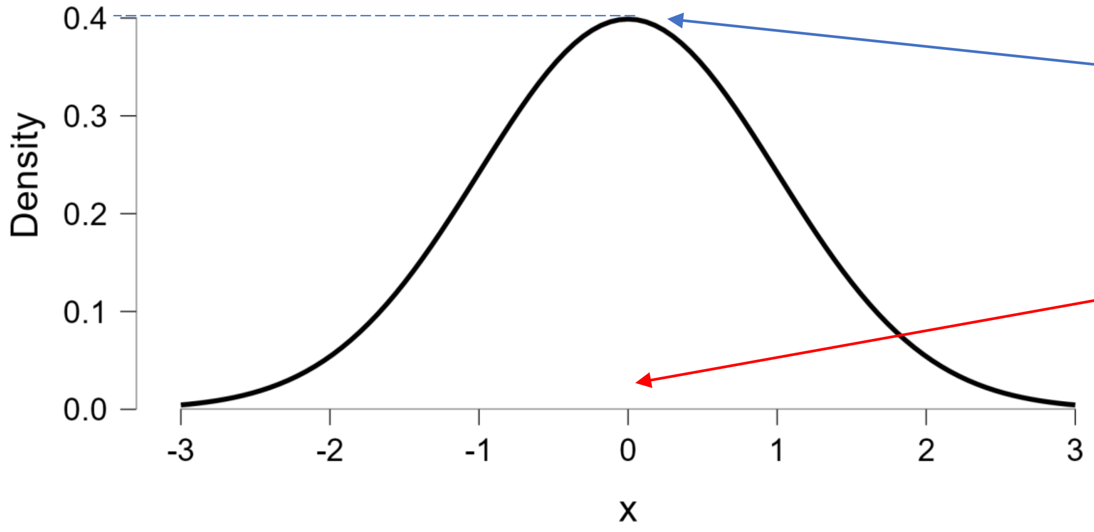
Στατιστικές δοκιμασίες που χρησιμοποιούνται επιπρόσθετα

Δοκιμασίες κανονικότητας

Οπτική δοκιμασία διάγραμμα QQ
Υπολογιστικές δοκιμασίες D'Agostino-Pearson, Anderson-Darling, Shapiro-Wilk and Kolmogorov-Smirnov


ΔΥΣΚΟΛΑ ΤΑ ΠΡΑΓΜΑΤΑ

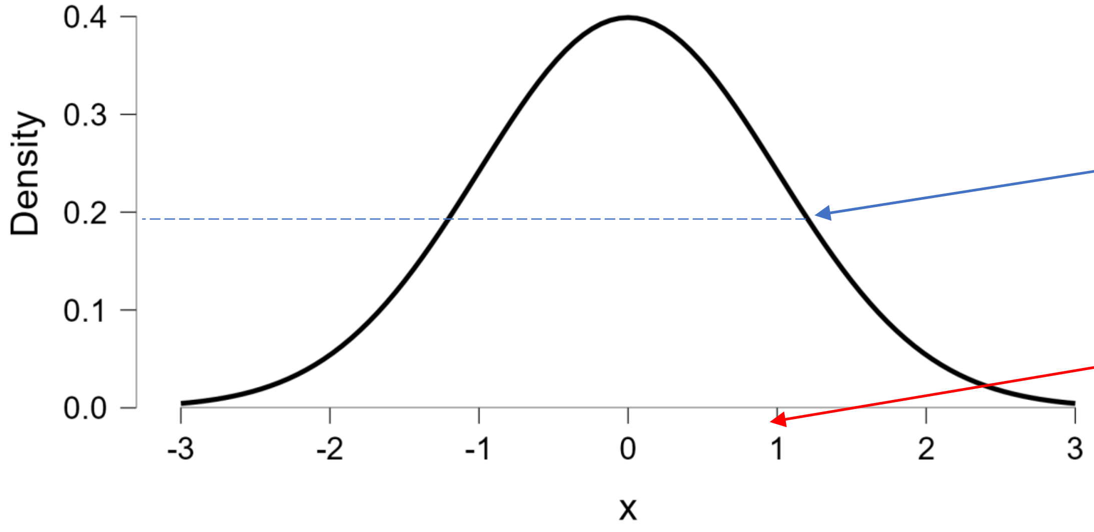
Έχει σχήμα καμπάνας



Μέγιστη πιθανότητα να συναντήσω την τιμή αυτή

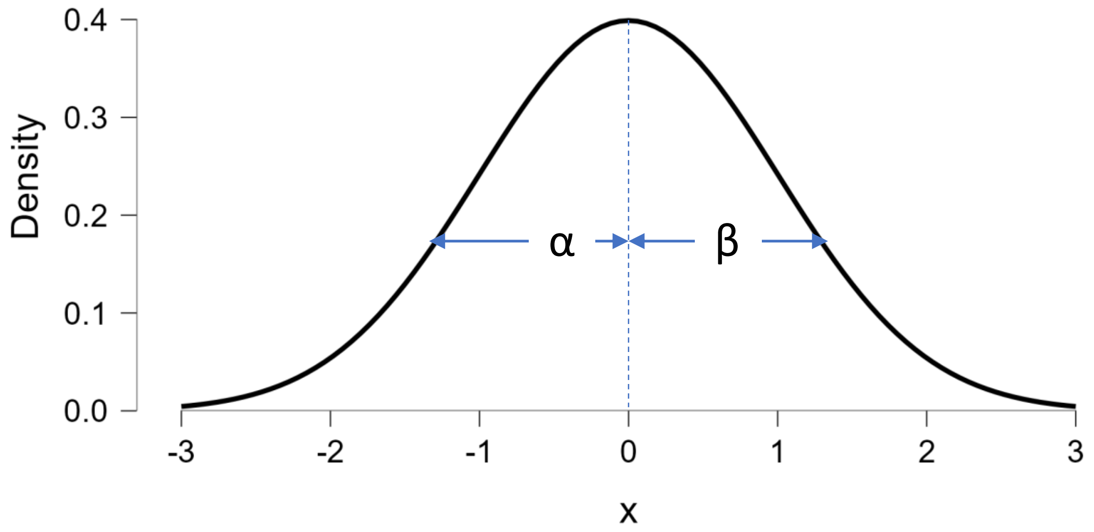
Οι πιθανότητες να συναντήσουμε διαφορετικές τιμές δεν είναι ίδια. Πιο πιθανό να συναντήσουμε τιμές κοντά στο κέντρο της κατανομής.

Η πιθανότητα με την οποία εμφανίζονται διάφορες τιμές δεν είναι γραμμική



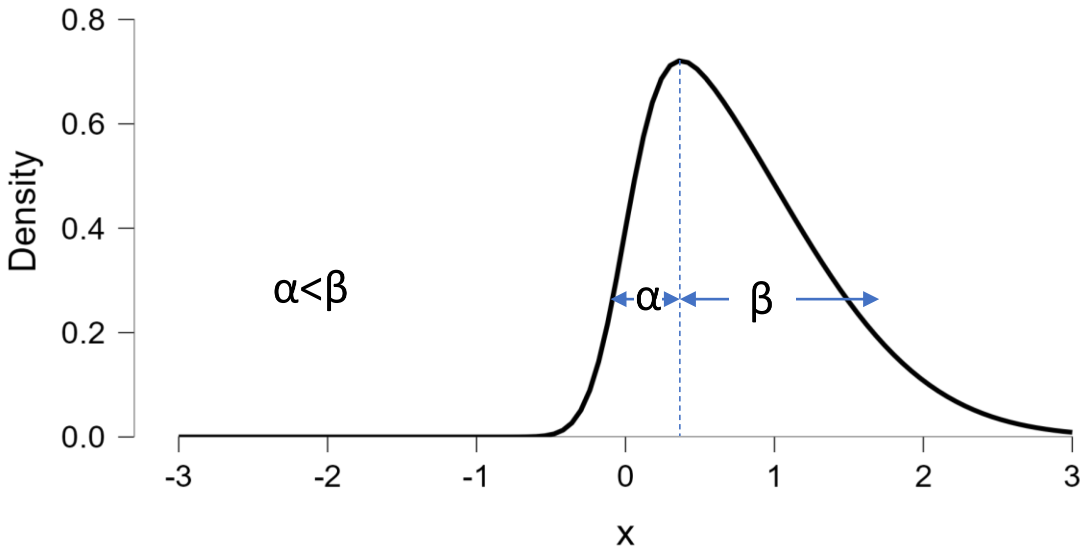
Μικρότερη πιθανότητα να συναντήσω την τιμή αυτή

Έχει συμμετρία και έλλειψη λοξότητας (skewness)

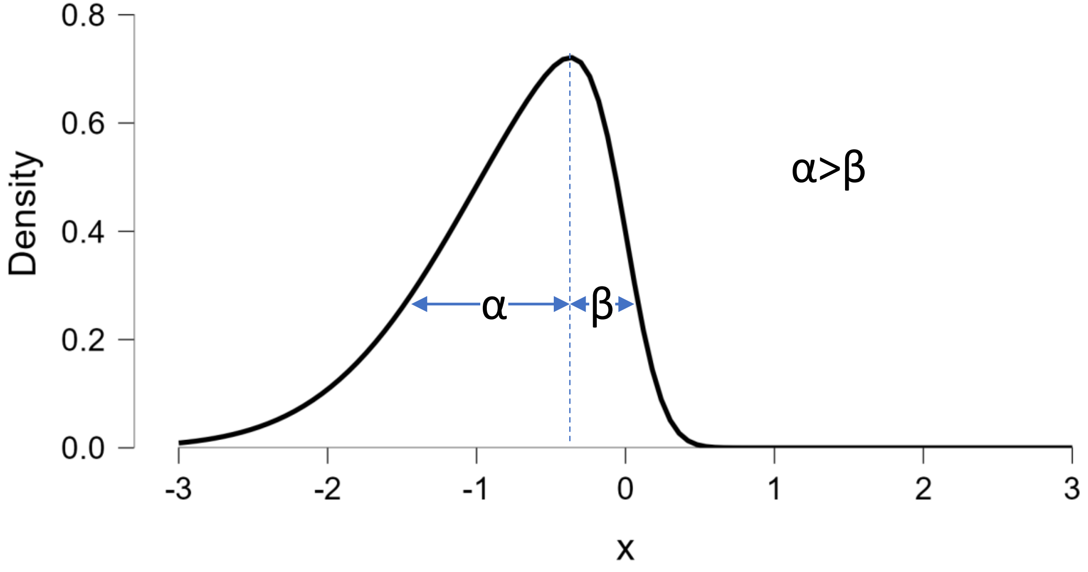


Συμμετρική κατανομή (προυπόθεση για να είναι κανονική)

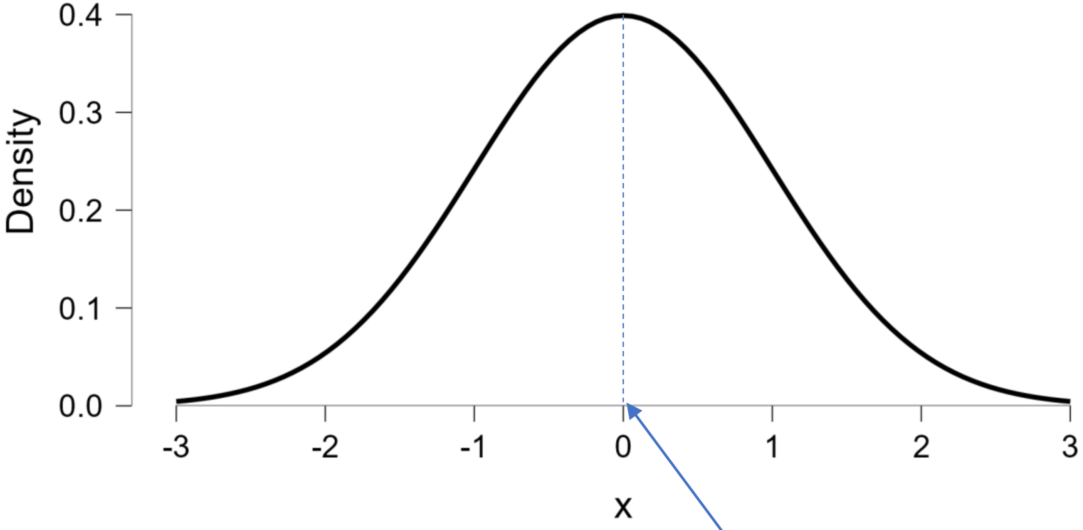
Συμμετρία α=β



Δεν είναι κανονικές κατανομές

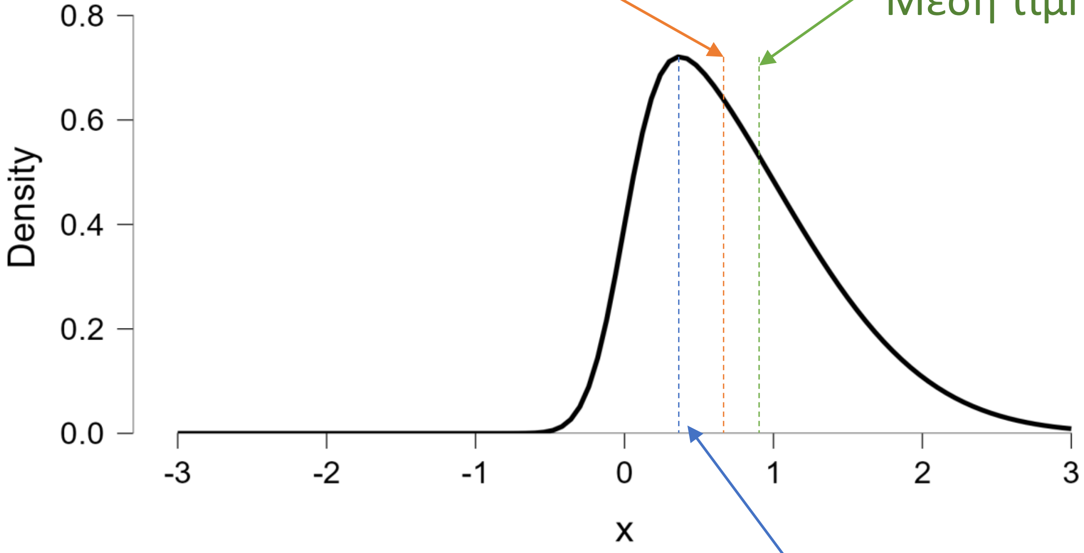


Η μέση τιμή συμπίπτει με τη διάμεση



Διάμεση τιμή

Μέση τιμή



Μέση τιμή
Διάμεση τιμή
Επικρατούσα τιμή

Συμπίπτουν

Επικρατούσα τιμή

διάμεση τιμή επικρατούσα τιμή

Μέση τιμή ή μέσος όρος (average ή mean value)
Σωστά ο αριθμητικός μέσος όρος

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

Αν όμως υπάρχει διαφορετική αξιολόγηση για την κάθε μέτρηση που συνεισφέρουν κατά διαφορετικό τρόπο τότε μπορούμε να χρησιμοποιήσουμε την μέση τιμή με διαφορετικό στατιστικό βάρος

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_n x_n}{w_1 + w_2 + w_3 + \cdots + w_n} = \frac{\sum x_i}{\sum w_n}$$

Διάμεση τιμή

Προβλήματα μέσης τιμής

1. Επηρεάζεται πολύ από τις έκτροπες τιμές

1, 1, 1, 1, 1 → 1

1, 1, 1, 1, 2 → 1,2 τιμή 20% μεγαλύτερη

2. Ο μέσος όρος μιας κανονικής κατανομής δεν λαμβάνει υπόψη το γεγονός ότι η πιθανότητα να συναντήσουμε τιμές όλο και μακρύτερα από το μέσο όρο μικραίνει. Στο μέσο όρο η απίθανη μακρινή τιμής έχει το ίδιο στατιστικό βάρος με την πολύ πιθανή τιμή που βρίσκεται κοντά στο μέσο όρο

Τι είναι η διάμεση τιμή

Η τιμή που χωρίζει τον πληθυσμό στα δύο αν οι τιμές είναι μονές ή ο μέσος όρος των δυο μεσαίων τιμών αν είναι ζυγές.

Τι πλεονέκτημα έχει? Ότι (προφανώς) αδιαφορεί για έκτροπες τιμές.

Ανήκει στα ανθεκτικά στατιστικά μεγέθη (robust)

Διάμεση τιμή

Προβλήματα μέσης τιμής

1. Επηρεάζεται πολύ από τις έκτροπες τιμές

2. Ο μέσος όρος μιας κανονικής κατανομής δεν λαμβάνει υπόψη το γεγονός ότι η πιθανότητα να συναντήσουμε τιμές όλο και μακρύτερα από το μέσο όρο μικραίνει. Στο μέσο όρο η απίθανη μακρινή τιμής έχει το ίδιο στατιστικό βάρος με την πολύ πιθανή τιμή που βρίσκεται κοντά στο μέσο όρο

1, 1, $\boxed{1,}$ 1, 1 → 1

1, 1, $\boxed{1,}$ 1, 2 → 1,2 τιμή 20% μεγαλύτερη

→ Η διάμεση τιμή παραμένει η ίδια

Τι είναι η διάμεση τιμή

Η τιμή που χωρίζει τον πληθυσμό στα δύο αν οι τιμές είναι μονές ή ο μέσος όρος των δυο μεσαίων τιμών αν είναι ζυγές.

Τι πλεονέκτημα έχει? Ότι (προφανώς) αδιαφορεί για έκτροπες τιμές.

Ανήκει στα ανθεκτικά στατιστικά μεγέθη (robust)

Διάμεση τιμή

Προβλήματα μέσης τιμής

1. Επηρεάζεται πολύ από τις έκτροπες τιμές

4, 5, 1, 1, 2

Επικρατέστερη τιμή είναι το 1 γιατί εμφανίζεται δυο φορές

2. Ο μέσος όρος μιας κανονικής κατανομής δεν λαμβάνει υπόψη το γεγονός ότι η πιθανότητα να συναντήσουμε τιμές όλο και μακρύτερα από το μέσο όρο μικραίνει. Στο μέσο όρο η απίθανη μακρινή τιμής έχει το ίδιο στατιστικό βάρος με την πολύ πιθανή τιμή που βρίσκεται κοντά στο μέσο όρο

Ποια άλλη τιμή είναι χρήσιμη? Η επικρατέστερη τιμή που είναι αυτή με την υψηλότερη συχνότητα εμφάνισης

Υπάρχουν άλλα στατιστικά μεγέθη σχετιζόμενα με τον μέσο όρο

Πολλά αλλά σπανίως χρησιμοποιούμενα

1. Ο γεωμετρικός μέσος όρος

$$\overline{x_g} = \sqrt[n]{x_1 x_2 x_3 \ldots x_n}$$

Χρησιμοποιείται για πολύ ασύμμετρες κατανομές

2. Ο αρμονικός μέσος όρος

Χρησιμοποιείται συχνά όταν θέλουμε να υπολογίσουμε αντίστροφες σχέσεις

$$\overline{x_h} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_1} + \frac{1}{x_1} + + \frac{1}{x_n}\right)} = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$$

Η έννοια της επαναληπτικότητας των μετρήσεων

Πολύ σημαντικό πρόβλημα της χημικής ανάλυσης αλλά και της επιστήμης των μετρήσεων είναι αν οι τιμές που λαμβάνονται παρουσιάζουν προβλέψιμη τυχαιότητα.

Δηλαδή κατά την επανάληψή τους τα αποτελέσματα δεν είναι χαοτικά αλλά ακολουθούν κάποιο στατιστικό μοντέλο

Έντουαρντ Λόρεντζ
Χάος: Όταν το παρόν καθορίζει το μέλλον, αλλά η προσέγγιση του παρόντος δεν προσδιορίζει κατά προσέγγιση το μέλλον.

Wikipedia

Η θεωρία του Χάους μελετά τη συμπεριφορά ορισμένων μη γραμμικών δυναμικών συστημάτων, που είναι ιδιαίτερα ευαίσθητα στις αρχικές συνθήκες, ένα αποτέλεσμα το οποίο ευρέως αναφέρεται ως το φαινόμενο της πεταλούδας. Μικρές διαφορές στις αρχικές συνθήκες (όπως αυτές που οφείλονται σε σφάλματα στρογγυλοποίησης σε αριθμητικούς υπολογισμούς) αποδίδουν πολύ διαφορετικά αποτελέσματα για τα δυναμικά συστήματα, καθιστώντας τη μακροπρόθεσμη πρόβλεψη αδύνατη σε γενικές γραμμές. Αυτό συμβαίνει παρ' όλο που αυτά τα συστήματα είναι αιτιοκρατικά ("ντετερμινιστικά"), πράγμα που σημαίνει ότι η μελλοντική συμπεριφορά τους καθορίζεται πλήρως από τις αρχικές συνθήκες τους, χωρίς να εμπλέκονται τυχαίες παράμετροι. Με άλλα λόγια, **η ντετερμινιστική φύση αυτών των συστημάτων δεν τα κάνει προβλέψιμα**

https://itp.uni-frankfurt.de/~gros/Vorlesungen/SO/simulation_example/
https://www.myphysicslab.com/pendulum/compare-pendulum-en.html

Ας ορίσουμε τα μεγέθη που μας περιγράφουν την επαναληψιμότητα

Είναι αρκετά και στο τέλος θα κρατήσουμε ένα

Μέτρα επαναληψιμότητας (επαναληπτικότητας?)

Το απλούστερο είναι το εύρος

Εύρος

Διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής

$$R = x_{max} - x_{min}$$

SOS ισχυει για ~10 δείγματα

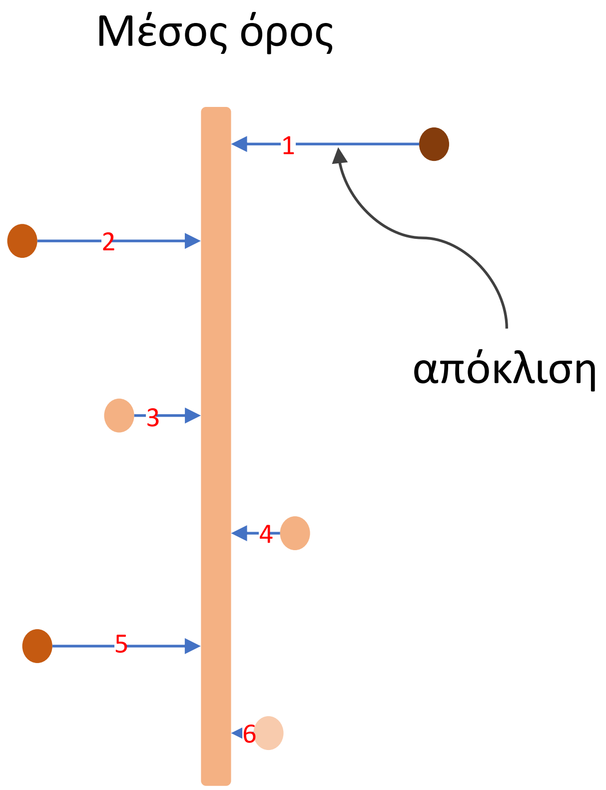
Είναι δυνατός ο υπολογισμός της απόκλισης (που θα δούμε στη συνέχεια) από το εύρος με τη βοήθεια της παρακάτω σχέσης

$$R\sigma_R = c \times R$$

Οι συντελεστές c δίνονται από σχετικό πίνακα

Μια ιδέα είναι

Να βρούμε την απόκλιση από τον μέσο όρο κάθε μέτρησης.
Αν το άθροισμα των αποκλίσεων είναι μεγάλο δεν επαναλαμβάνεται και πολύ το πείραμα

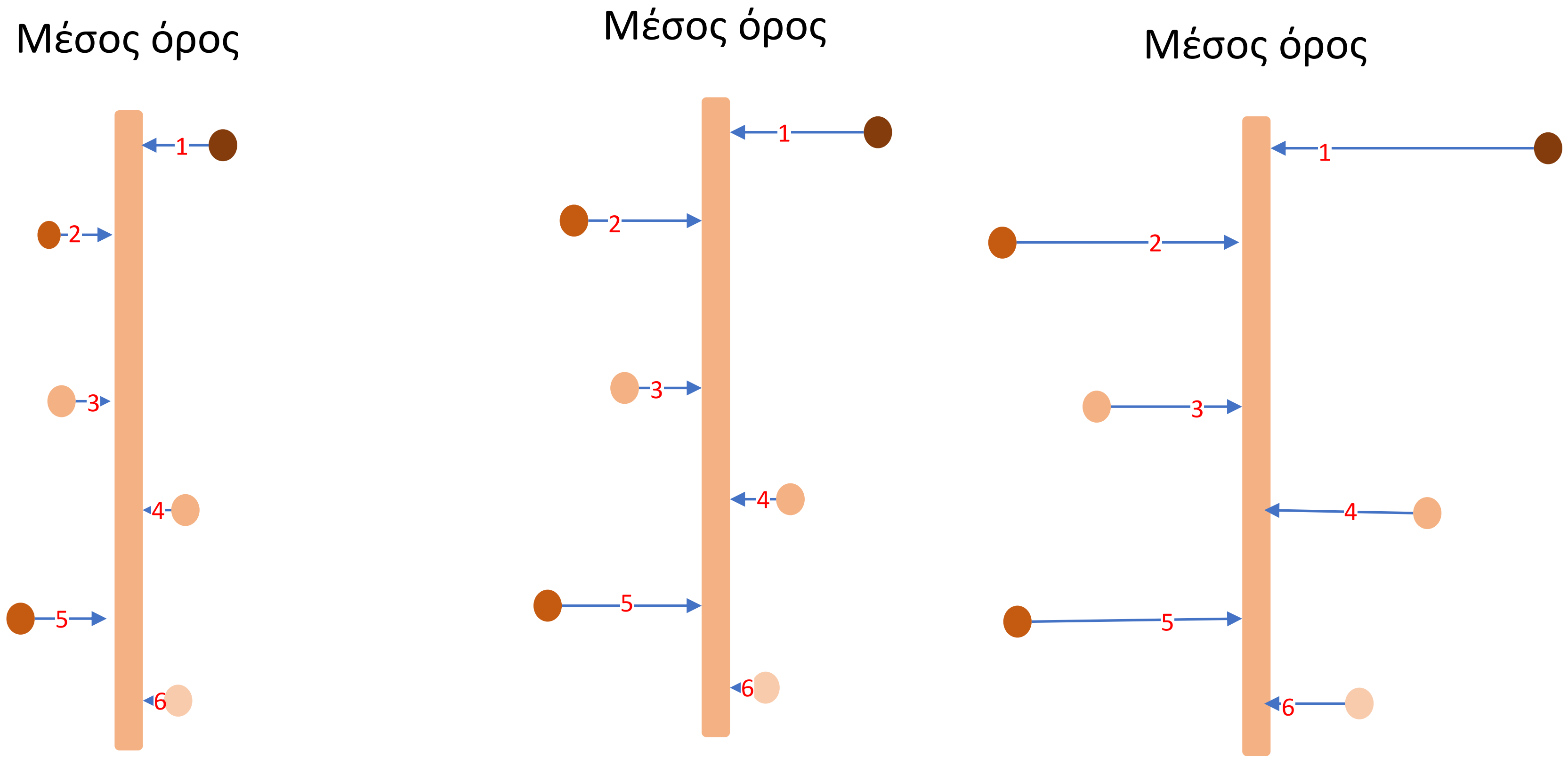


1+2+3+4+5+6

Το άθροισμα των αποκλίσεων

Μια ιδέα είναι

Να βρούμε την απόκλιση από τον μέσο όρο κάθε μέτρησης.
Αν το άθροισμα των αποκλίσεων είναι μεγάλο δεν επαναλαμβάνεται και πολύ το πείραμα



Μέση απόκλιση

Θα μπορούσε η απόκλιση αντί να τετραγωνίζεται εναλλακτικά να παρέχεται ως απόλυτη τιμή. Έτσι ορίζουμε την μέση απόκλιση $\bar{d}$

$$\bar{d} = \frac{|\bar{x} - x_i|}{n} = \frac{|\bar{x} - x_i| + |\bar{x} - x_i| + |\bar{x} - x_i| + \cdots + |\bar{x} - x_i|}{n}$$

Πρόβλημα της μέσης απόκλισης
Δίνει την ίδια σημασία σε όλες της αποκλίσεις ενώ στην κανονική κατανομή αυτές που βρίσκονται κοντά στο κέντρο είναι πιθανότερες

Τυπική απόκλιση

Θα μπορούσε η απόκλιση να τετραγωνίζεται ώστε πάντα να έχει θετικό πρόσημο (όταν παρατηρείται λάθος θα θέλαμε πάντα να αθροίζεται). Ορίζουμε την μέση (ΔΕΙΓΜΑΤΙΚΗ) απόκλιση s

$$s = \sqrt{\frac{\sum(\bar{x} - x_i)^2}{n - 1}}$$

Αν $s \to \infty$ τότε ο μέσος όρος γίνεται $\bar{x} \to \mu$ (ο πληθυσμιακός μέσος όρος) και το $s \to \sigma$ (η πληθυσμιακή τυπική απόκλιση)

$$\sigma = \sqrt{\frac{\sum(\mu - x_i)^2}{n}}$$

Οι τρεις διαφορές

Οι δύο είναι αναμενόμενες σ και μ
Η τρίτη n→n-1 οφείλεται στην αμελητέα συμμετοχή του -1 σε άπειρο πλήθος παρατηρήσεων

$$\sigma = \sqrt{\frac{\sum(\mu - x_i)^2}{n}} \qquad s = \sqrt{\frac{\sum(\bar{x} - x_i)^2}{n-1}}$$

Τι είναι το n-1 όμως

Οι **βαθμοί ελευθερίας** (θα δούμε παρακάτω)

$$\sigma = \sqrt{\frac{\sum(\mu - x_i)^2}{n}}$$

ΠΛΗΘΥΣΜΙΑΚΗ ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ

$$s = \sqrt{\frac{\sum(\bar{x} - x_i)^2}{n-1}}$$

ΔΕΙΓΜΑΤΙΚΗ ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ

SOS

Μια πιο εύκολη σχέση υπολογισμού της τυπικής απόκλισης

$$s = \sqrt{\frac{1}{n-1}\left[\sum x_i^2 - \frac{\sum x_i^2}{n}\right]}$$

Στη σχέση αυτή συμμετέχουν μόνο οι καταχωρίσεις μιας σειράς αριθμών (δηλαδή τιμές δειγμάτων) και όχι το $\bar{x}$ που περιπλέκει τα πράγματα καθώς πρέπει να υπολογίζεται και αυτό κάθε φορά που μεταβάλλονται οι καταχωρίσεις

Τι γίνεται για δείγματα διαφορετικών πληθυσμών μιας μεθόδου

$$s = \sqrt{\frac{\sum\left(\overline{x_1} - x_{i,1}\right)^2 + \left(\overline{x_2} - x_{i,2}\right)^2 + \left(\overline{x_3} - x_{i,3}\right)^2 + \cdots + \left(\overline{x_k} - x_{i,k}\right)^2}{n-k}}$$

ή πιο συνεπτυγμένα

$$s = \sqrt{\frac{\sum\sum(x_j - x_{j,i})^2}{n-k}}$$

$x_{i,1}$ τα δείγματα του πρώτου υποπληθυσμού που χρησιμοποιούμε στη μέθοδο και $\overline{x_1}$ ο μέσος όρος του πρώτου υποπληθυσμού κλπ

n ο συνολικός αριθμός των μετρήσεων και

k ο συνολικός αριθμός των δειγμάτων


%RSD

% σχετική τυπική απόκλιση (% relative standard deviation)

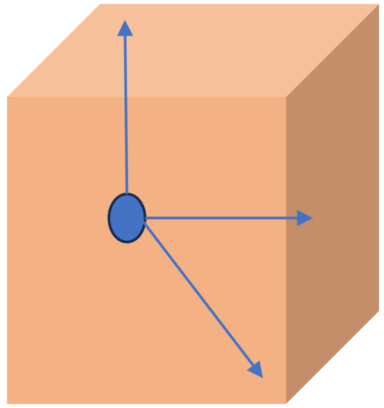
Ίσως το πιο χρησιμοποιούμενο μέγεθος

Τι μας λέει? Πόσο % είναι μεγαλύτερη η τυπική απόκλιση από τον μέσο όρο

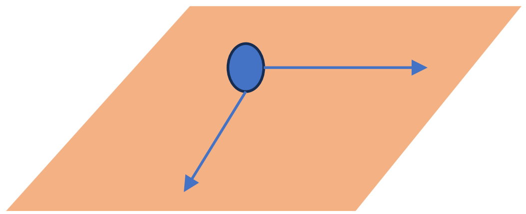
$$\% \, s_{rel} = \frac{s}{\bar{x}} \times 100$$

Βαθμοί ελευθερίας (ένα δύσκολα κατανοήσιμο διαισθητικό μέγεθος)

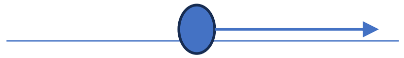
Ας δούμε ένα παράδειγμα από τη φυσική



3 βαθμοί ελευθερίας



2 βαθμοί ελευθερίας



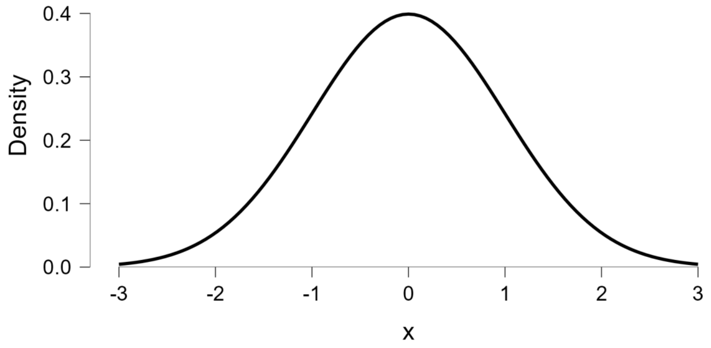
1 βαθμός ελευθερίας

Μια ευθεία έχει n-2 βαθμούς ελευθερίας γιατί δύο τουλάχιστον σημεία της είναι καθορισμένα

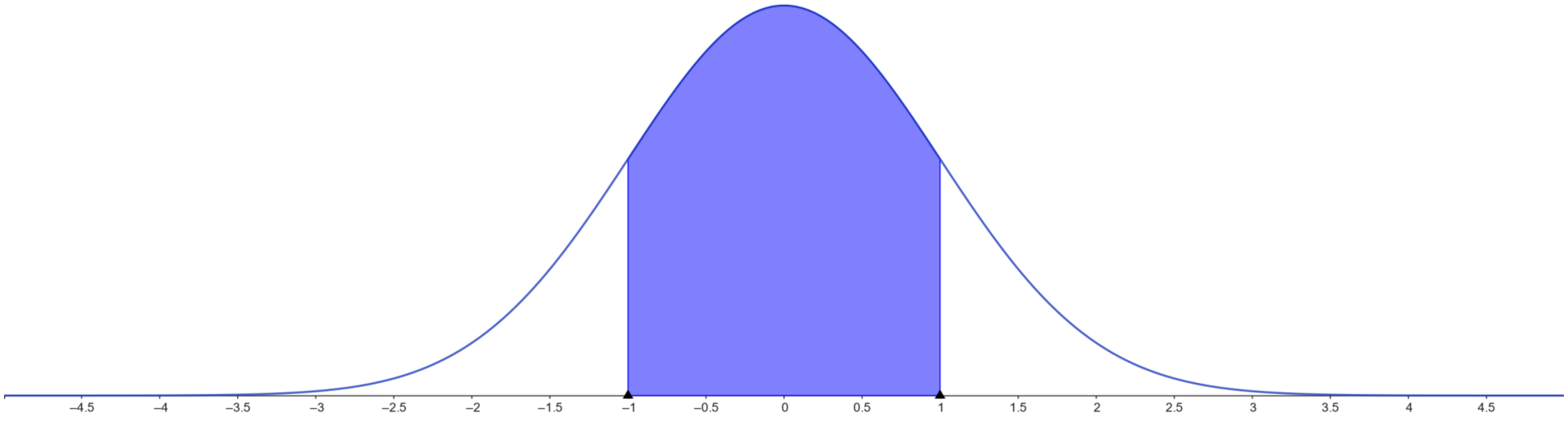
Μια κατανομή με πεπερασμένα σημεία? N-1 πχ τρεις μετρήσεις $x_1 + x_2 + x_3 = \bar{x}$ και ξέρουμε τις δύο πρώτες η Τρίτη είναι αυστηρά καθορισμένη

Επιστρέφουμε στην κατανομή Gauss

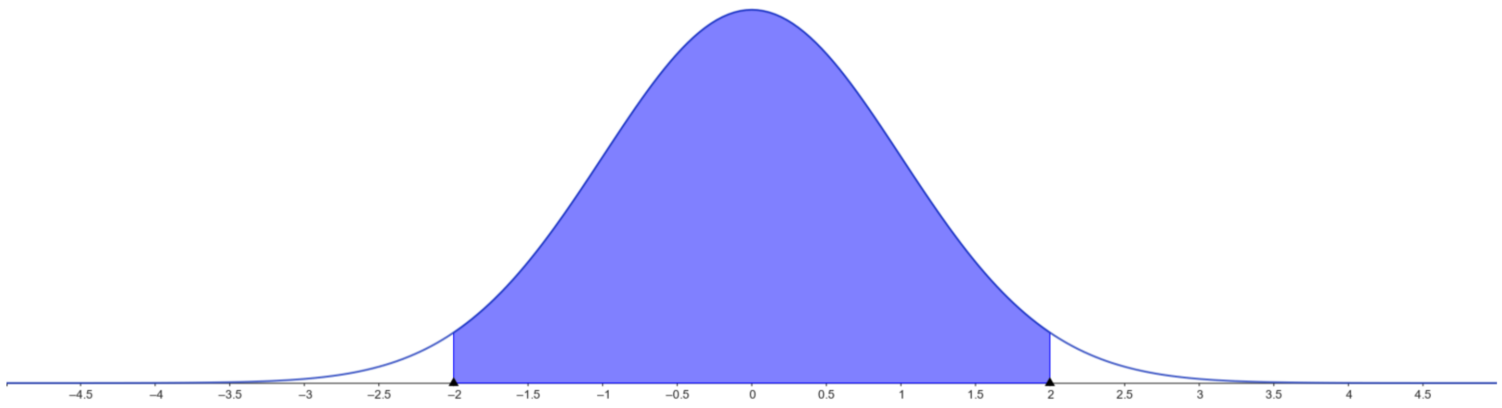
Η κατανομή Gauss έχει την εξαιρετική δυνατότητα να είναι καθοριζόμενη από την τυπική απόκλιση



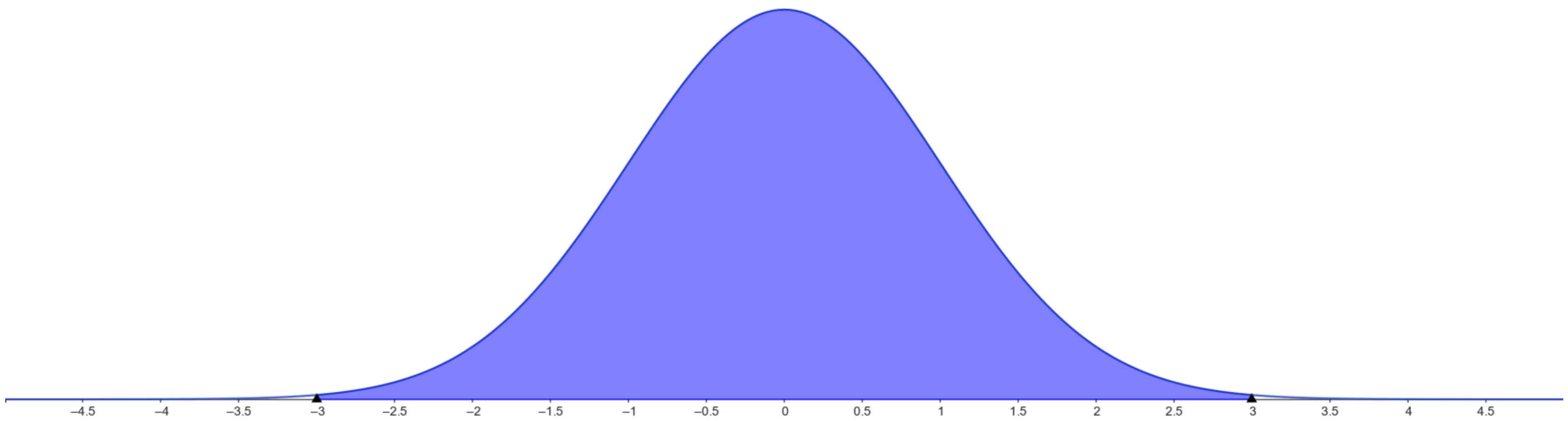
Πως? Αν κινηθούμε **μια** τυπική απόκλιση από τον μέσο όρο αριστερά ή δεξιά τότε ξέρουμε ότι 68 στις 100 φορές που θα κάνουμε μια μέτρηση ο μέσος όρος θα βρίσκεται στην περιοχή αυτή



Δύο τυπικές αποκλίσεις τότε 95.55%



τρεις τυπικές αποκλίσεις τότε 99.73%



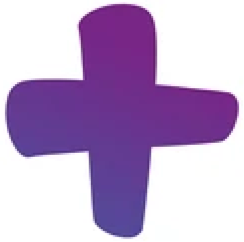
Δηλαδή ξέρουμε με σιγουριά ο μέσος όρος βρίσκεται στο όριο των τριών τυπικών αποκλίσεων

Αυτός είναι και ο κανόνας 68-96-99.7 (που αναφέρεται σε τυπικές αποκλίσεις)

Τι σημαίνει τελικά η κανονική κατανομή?

Είναι η έκφραση της τυχαιότητας. Όταν κάτι είναι τυχαίο λεμέ ότι η πιθανότητα να το παρατηρήσουμε ακολουθεί κανονική κατανομή και συνεπώς και τα σφάλματα του $\bar{x} - x_i$ ακολουθούν και αυτά κανονική κατανομή.
Αλλιώς το ίδιο
Υπάρχει διασπορά τιμών γύρω από την αληθή τιμή. Η διασπορά των τιμών ακολουθεί κατανομή κατά Gauss.

Κατανομή κατά Gauss η κανονική κατανομή. Δεν ειναι πολύ σαφές. Υπάρχει η κατανομή που έχει $\bar{x} = 0$ και $\sigma^2 = 1$. Συχνά ονομάζεται κανονικοποιημένη κατανομή Gauss.

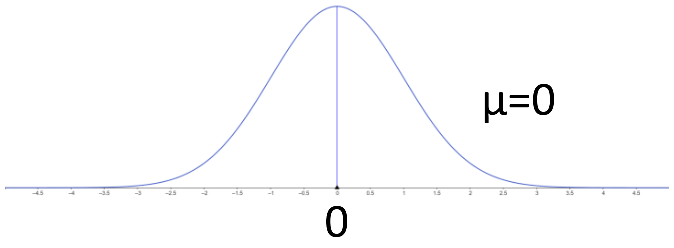
   

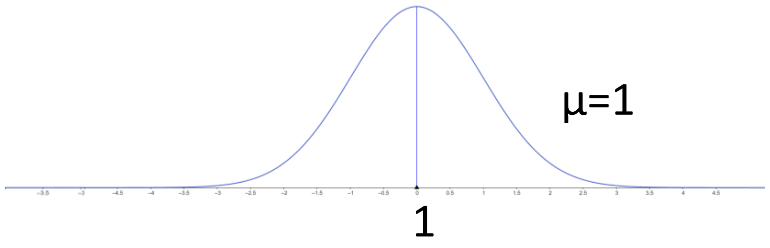
Εξίσωση της κατανομής Gauss

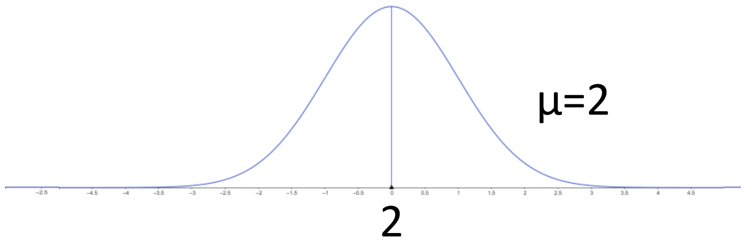
$$P_x = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

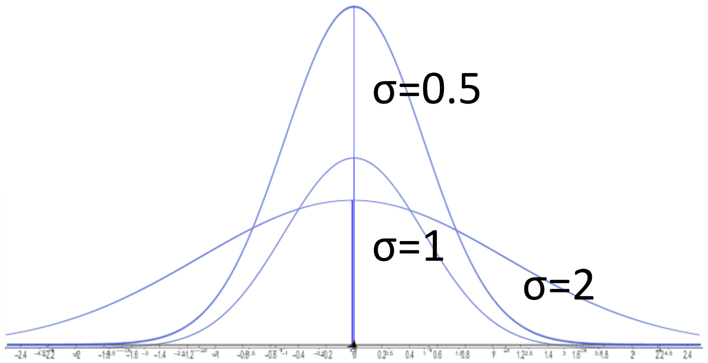
Όπου
μ η αληθής τιμή του πληθυσμού
σ ο πληθυσμιακή τυπική απόκλιση
και x η εκάστοτε τιμή

Τι βλέπουμε? Ότι η κατανομή εξαρτάται από δυο όρους το μ και το σ
Το μ καθορίζει τη θέση της κατανομής ενώ το μ το πλάτος της









Πως υπολογίζεται η πιθανότητα μια τιμή να βρίσκεται μεταξύ κάποιων ορίων (intervals)?

Από πίνακες (στην επόμενη σελίδα) ή με τη βοήθεια υπολογιστή (

ΑΛΛΑ

Πρώτα πρέπει να κανονικοποιήσουμε τις τιμές της κατανομής. Πως γίνεται αυτό?

Με τις τιμές z

Πως τις υπολογίζουμε? Απλά

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

Δηλαδή πόσες φορές η απόσταση από τη μέση τιμή είναι μεγαλύτερη από την τυπική απόκλιση?

Γιατί χρησιμοποιούμε αυτή την κανονικοποίηση?

Γιατί οι κατανομές έχουν άπειρο πλήθος τιμών μ και σ. Θα επρεπε να υπολογίσουμε έτσι άπειρο πλήθος κατανομών. Με την z κανονικοποίηση, όλες οι κατανομές λύνονται καθώς όλες αποκτουν μ=0 και z=1

Παράδειγμα

Επαναλαμβανόμενες μετρήσεις μια χημικής ανάλυσης μας δίνουν μέσο όρο 10.15mL με τυπική απόκλιση 0.02 mL. Πόσο ποσοστό των μετρήσεων θα βρίσκεται μεταξύ 10.12 mL και 10.20 mL?

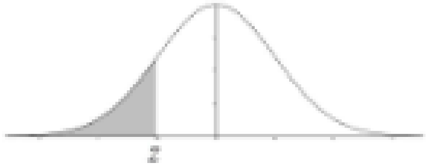
Z=(10.12-10.15)/0.02=-1.5 → πίνακας επόμενης σελίδας F(-1.5) =0.0668 *η ΣΥΝΟΛΙΚΗ πιθανότητα να βρεθεί μια μέτρηση 10.12mL*

Z=(10.20-10.15)/0.02=-2.5 → πίνακας επόμενης σελίδας F(2.5) =0.9938 *η ΣΥΝΟΛΙΚΗ πιθανότητα να βρεθεί μια μέτρηση 10.20mL*

Διαφορά=0.927 *η ΣΥΝΟΛΙΚΗ πιθανότητα να βρεθεί μια μέτρηση10.12mL και 10.20 mL*

Standard Normal Cumulative Probability Table for Negative Values of z

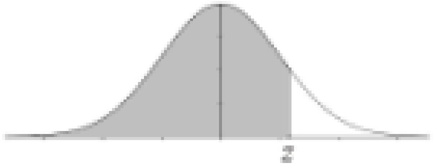
The probabilities in this table represent $P(Z < z)$, the proportion of data less than z.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Standard Normal Cumulative Probability Table for Positive Values of z

The probabilities in this table represent $P(Z < z)$, the proportion of data less than z.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ

Σε απλά λόγια ο μέσος όρος των μέσων όρων πολλών μετρήσεων είναι ακριβέστερος από τον μέσο όρο όλων των μετρήσεων

Δεν φαινεται απλό αλλά ως παράδειγμα είναι αλούστερο

Αν έχουμε 50 μετρήσεις τότε έχουμε ένα μέσο όρο $\bar{x}$ που πλησιάζει την πραγματική τιμή μ

Αν χωρίσουμε τις μετρήσεις σε πεντάδες (10 πεντάδες προφανώς) και υπολογίσουμε τον μέσο όρο κάθε πεντάδας προκύπτουν δέκα αριθμοί. Ο μέσος όρος αυτών των δέκα αριθμών είναι πλησιέστερος στην πραγματική τιμή μ

Δηλαδή προτιμούμε αντί για τον μέσο όρο ΟΟΟΟΟΛΩΝ των παρατηρήσεων να χωρίσουμε τις παρατηρήσεις πχ σε πεντάδες, να καταγράψουμε τον μέσο όρο της κάθε πεντάδας και μετά να βγάλουμε τον μέσο όρων των καταγεγραμμένων μέσων όρων.

ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ

Σε απλά λόγια ο μέσος όρος των μέσων όρων πολλών μετρήσεων είναι ακριβέστερος από τον μέσο όρο όλων των μετρήσεων

Δεν φαινεται απλό αλλά ως παράδειγμα είναι αλούστερο

Αν έχουμε 50 μετρήσεις τότε έχουμε ένα μέσο όρο $\bar{x}$ που πλησιάζει την πραγματική τιμή μ

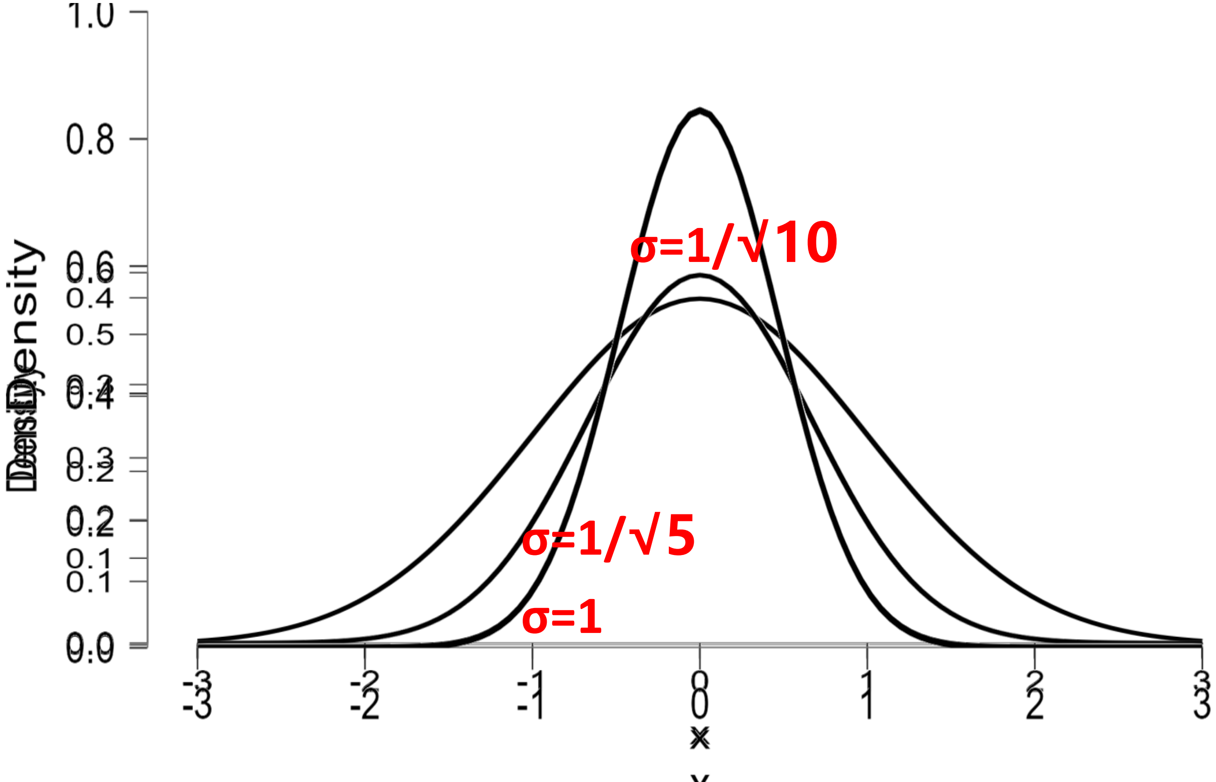
ες προφανώς) και υπολογίσουμε τον μέσο όρο κάθε πεντάδας προκύπτουν
ν είναι πλησιέστερος στην πραγματική τιμή μ

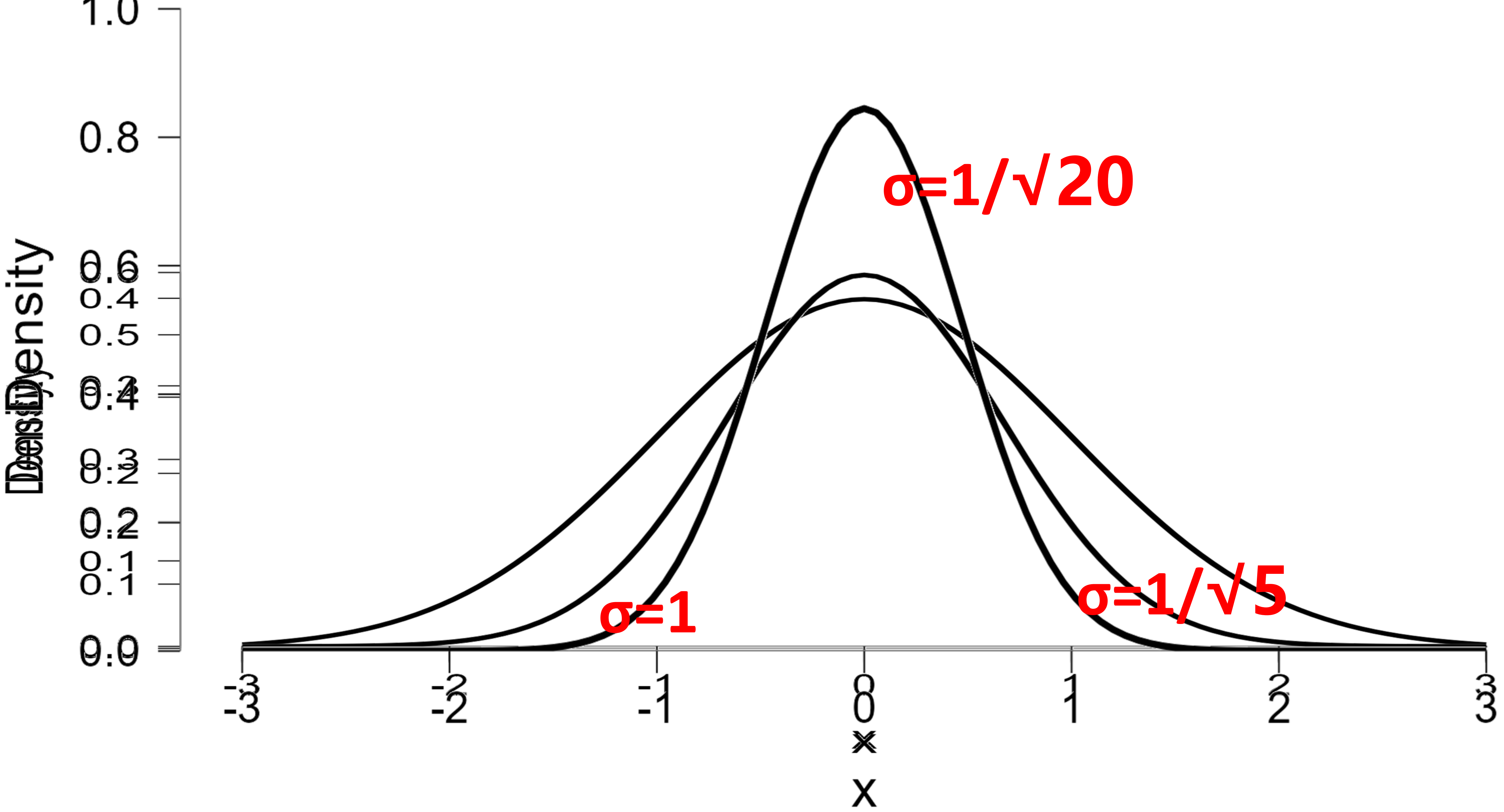
ΩΝ των παρατηρήσεων να χωρίσουμε τις παρατηρήσεις πχ σε πεντάδες, να
μετά να βγάλουμε τον μέσο όρων των καταγεγραμμένων μέσων όρων.

Ο μέσος όρος της κατανομής παραμένει ό ίδιος θέλοντας να προσεγγίσει την πραγματική τιμή μ
ΤΙ αλλάζει? Η τυπική απόκλιση
Ποια είναι η τυπική απόκλιση των μέσων όρων σε μια διαδικασία που εμπλέκει την κατανομή του μέσου όρου των δειγμάτων? Το τυπικό σφάλμα του μέσου όρου ή αλλιώς standard error of the mean (SEM)

$$SEM = \frac{\sigma}{\sqrt{n}}$$

Έστω μια κανονική κατανομή με μέσος όρο μ=0 και τυπική απόκλιση σ=1 από 50 δείγματα
Στο σχήμα φαίνεται η κατανομή όταν υπολογίσουμε τα δείγματα σε πεντάδες και δεκάδες μαζί με τον αρχικό πληθυσμό. Βλέπουμε ότι ο μέσος όρος είναι ίδιος και στις τρεις περιπτώσεις αλλά η τυπική απόκλιση μικραίνει. Μας βοηθά δηλαδή να υπολογίζουμε τους μέσους όρους




σ=1/√20
σ=1/√5
σ=1
Density
x

Έχουμε δει ότι μεταξύ 2σ και -2σ από τον μέσο όρο μ βρίσκονται το 95.55% των τιμών όταν κάνουμε ένα πείραμα. Είναι ενδιαφέρον ότι έχουμε επιλέξει να μιλάμε για το 95% (όχι το 95.55%) γιατί είναι στρογγυλός αριθμός. Αυτό μεταφραζεται σε 19/20 τιμές. Αν το υπολογίσουμε είναι

1.96σ

Άρα η υπολογιζόμενη τιμή του πληθυσμού από το δείγμα $\bar{x}$ έχει διαφορά $\bar{x} - \mu$ από την πραγματική του πληθυσμού.

Το $\bar{x} - \mu$ είναι το **σφάλμα** με το οποίο υπολογίζουμε την πραγματική διαφορά του πληθυσμού.

Όπως έχουμε πει το σφάλμα είναι **κανονικά κατανεμημένο**.

Άρα το σφάλμα ακολουθεί κανονική κατανομή, οπότε στο 95 % των περιπτώσεων (από -1,96σ ως +1,96σ) θα περιέχεται μεταξύ των δύο αυτών ορίων

Ας το γράψουμε – το σφάλμα 95 στις 100 φορές θα περιέχεται μεταξύ -1.96 σ και +1,96 σ

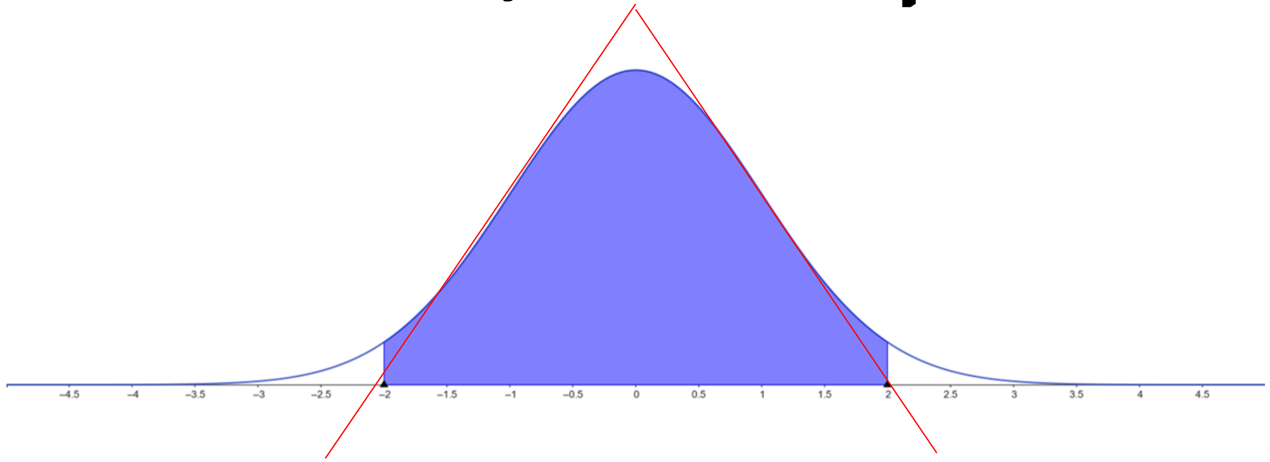
$$\mathbf{-1{,}96\sigma < \mu - \bar{x} < +1.96\sigma}$$

Αν χρησιμοποιήσουμε τους μέσους όρους του κεντρικού οριακού θεωρήματος

$$\mathbf{-1{,}96\sigma\sqrt{}n < \mu - \bar{x} < +1.96\sigma\sqrt{}n}$$

Προσθέτουμε $\bar{x}$ και στους τρεις όρους

$$\mathbf{\bar{x} -1{,}96\sigma\sqrt{}n < \mu - \bar{x} + \bar{x} < \bar{x} + 1.96\sigma\sqrt{}n \Rightarrow \bar{x} - 1{,}96\sigma\sqrt{}n < \mu < \bar{x} + 1.96\sigma\sqrt{}n}$$



Ας το σημειώσουμε λίγο διαφορετικά

$$\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

Όπου	z=1.96 για επίπεδο εμπιστοσύνης 95%
z=2.58 για επίπεδο εμπιστοσύνης 99%
z=2.97 για επίπεδο εμπιστοσύνης 99.7%

Αυτά ισχύουν για μεγάλο αριθμό δειγμάτων.

Τι κάνουμε όμως αν έχουμε πεπερασμένο αριθμό δειγμάτων

????

ΚΑΤΑΝΟΜΗ t
(κατανομή του Student)

Αν έχουμε μικρό αριθμό δειγμάτων η κατανομή Gauss δεν ισχύει. Αυτό το έδειξε ένα ντροπαλός μαθηματικός ο Gosset που υπέγραψε με το όνομα Student.

Ποιο είναι το βασικό γνώρισμα μιας κατανομής Student?

Ο πεπερασμένος πληθυσμός.

Πρόβλημα που οδήγησε στη δημιουργία της κατανομής. Έστω ότι έχω ένα κανονικό πληθυσμό. Μπορώ όμως να κάνω λίγα πειράματα (π.χ. 5) και από αυτά να βγάλω συμπερασμα για την τιμή μ του πλήθυσμού.

Τι σημαίνει αυτό? Σημαίνει ότι δεν μπορώ να υπολογίσω το μ (γιατί πρέπει να μετρήσω όλο τον πληθυσμο όπως έχουμε δει) αλλά δεν μπορώ να υπολογίσω και το σ (γιατί θα πρέπει να κάνω αρκετές πχ. 5αδες πειραμάτων για να έχω μια καλή εκτίμηση αλλά εχω μόνο 5). Τι έχω τελικά στη διάθεση μου? Το $\bar{x}$ και το s (για πεπερασμένο ΜΙΚΡΟ αριθμό δειγμάτων n)

Η κατανομή Student ισχύει με τις ίδιες προυποθέσεις όπως και η κανονική κατανομή. Βασικά είναι μια κανονική κατανομή σε μια πιο γενική περίπτωση (γιατί ισχύει και για μικρά και για μεγάλα δείγματα)

Τύπος της κατανομής Student

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Τι βλέπουμε?

Τύπος της κατανομής Student

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

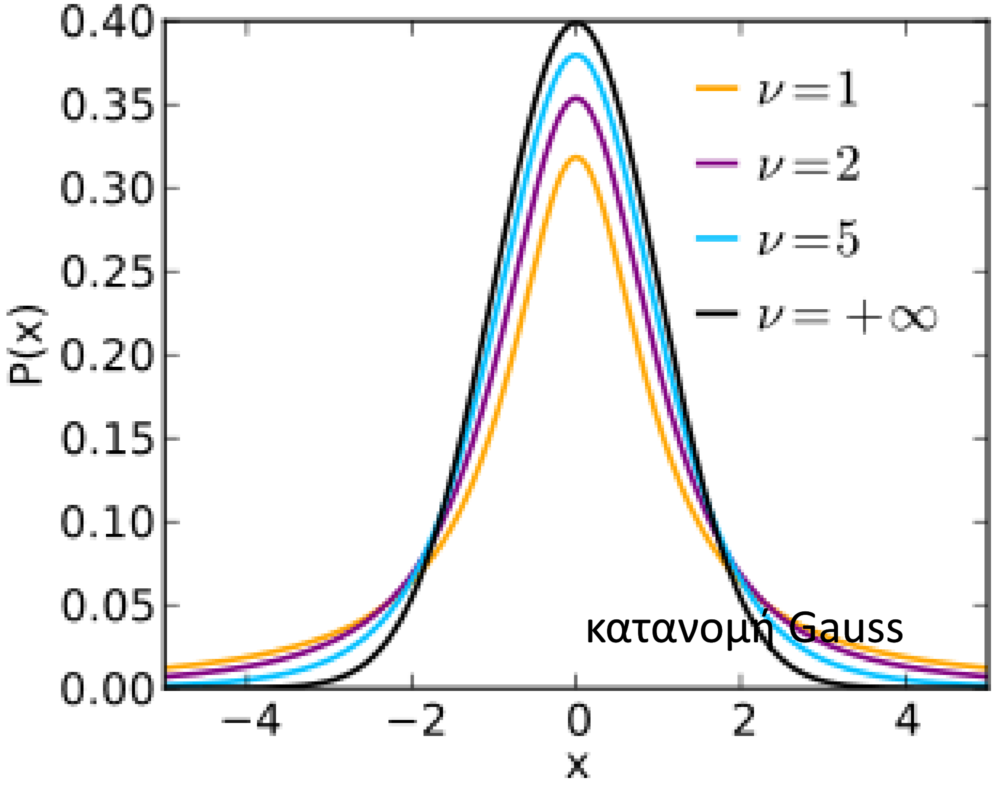
Τι βλέπουμε?

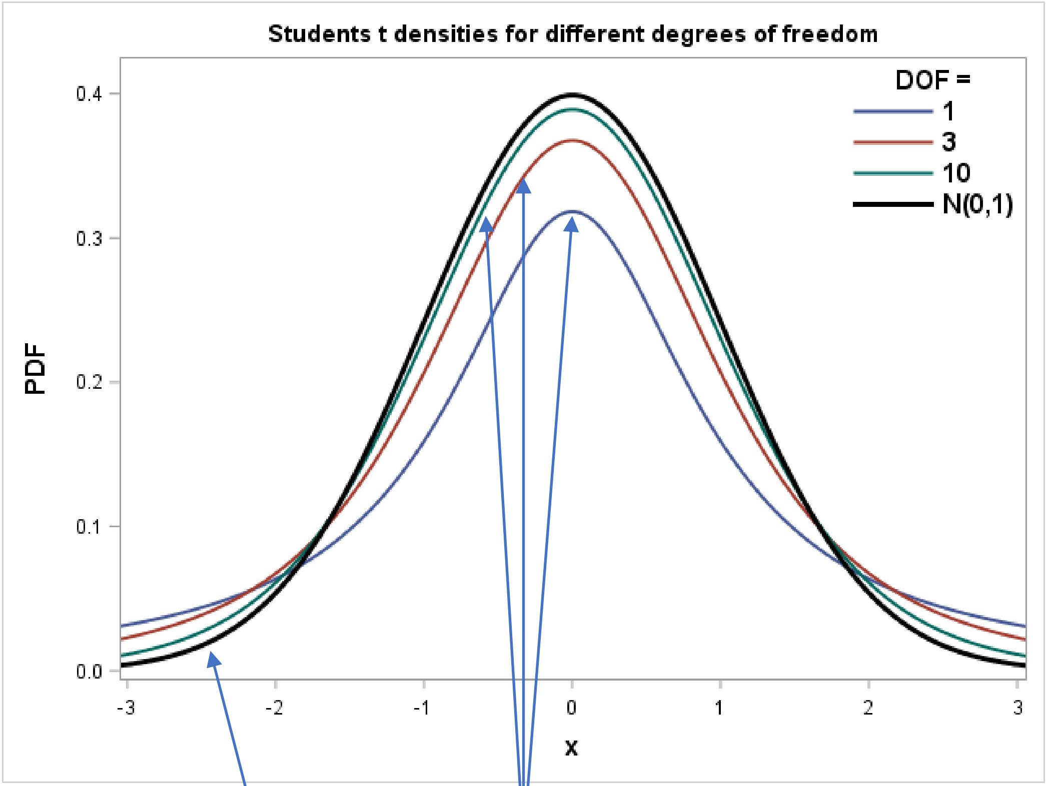
Τύπος της κατανομής Student

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Εξαρτάται μόνο από το ν! (των αριθμό των δειγμάτων) χμμμμ όχι ακριβώς ν=n-1 (βαθμοί ελευθερίας δηλαδή)

Τι είναι το Γ ? Είναι το (ν-1)! (ονομάζεται συνάρτηση Γ)
Τι είναι το t? Μια τυχαία μεταβλητή



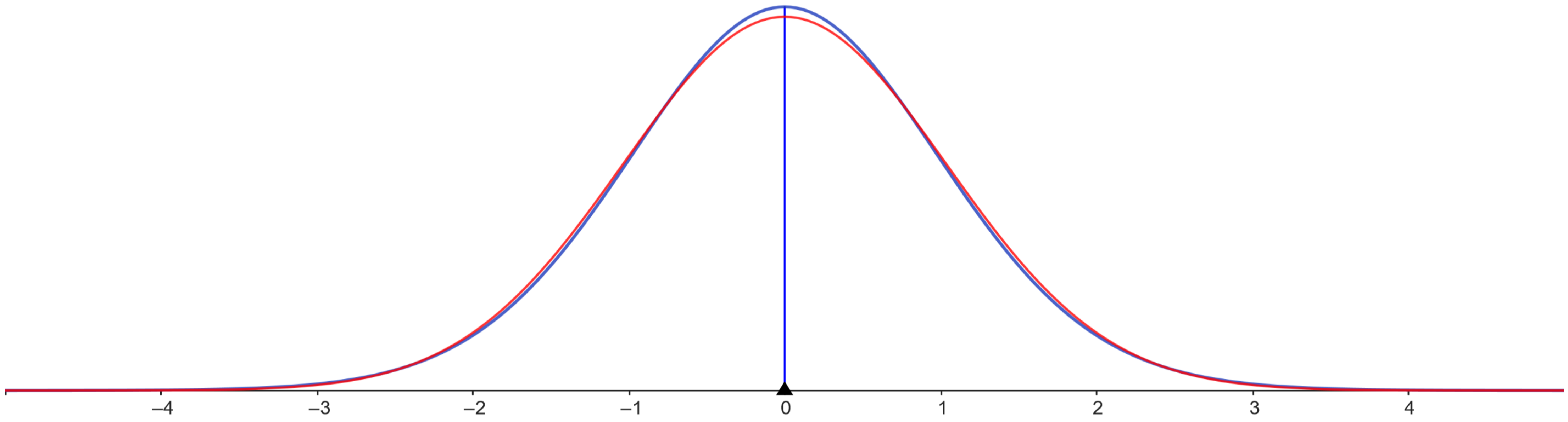


Κατανομές Student

κατανομή Gauss

Είναι διαφορετική για κάθε βαθμό ελευθερίας! Στην αρχή v=1 είναι πλατιά στα άκρα γιατί δεν ξέρουμε με τόσο λίγα δείγματα (2) την τυπική της απόκλιση. Όταν απειρίζεται σχεδόν ακουμπά στον άξονα των χ μετά από 3 τυπικές αποκλίσεις (99.7…). Που το έχουμε δει αυτό? ΤΙ σημαίνει? Για άπειρα δείγματα είναι η γνωστή μας κατανομή Gauss

Περίπου στα 30 δείγματα οι δύο κατανομές είναι πολύ παρόμοιες μεταξύ τους



Μετά από 30 δείγματα θεωρούμε ότι έχουμε πια κανονική κατανομή

Πως χρησιμοποιούμε την κατανομή Student?

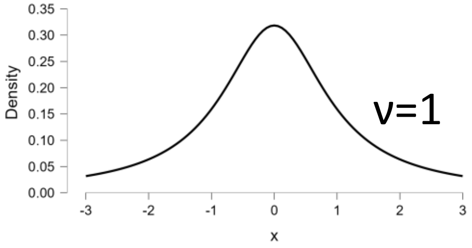
Είδαμε προηγουμένως ότι

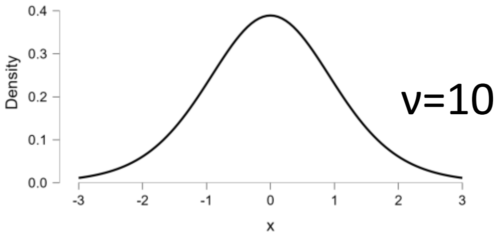
$$\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

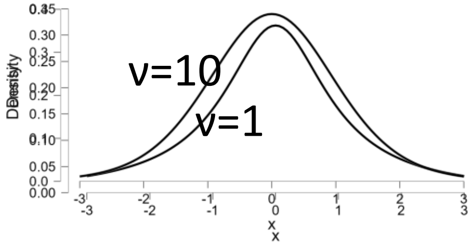
Επειδή όμως δεν ξέρουμε πλέον το σ (το σ από 5 μόνο δείγματα είναι πολύ κακή εκτιμήτρια συνάρτηση) θα χρησιμοποιήσουμε τη δειγματική εκτιμήτρια s

$$\mu = \bar{x} \pm t\frac{s}{\sqrt{v}}$$

Τώρα στον παρονομαστή βρίσκεται το ν. Τι είναι το ν? Ονομάζεται βαθμοί ελευθερίας και σχετίζεται με τον αριθμό δειγμάτων με τη σχέση ν=n-1 στην περίπτωση του στατιστικού t.
Και το t? Είναι αντίστοιχο του z. Είχαμε δει στην κατανομή Gauss πως το z είχε συγκεκριμένες τιμές σταθερές για κάθε επίπεδο εμπιστοσύνης, πχ 1.96 για επίπεδο εμπιστοσύνης 95%. Το t δεν έχει αυτή την πολυτέλεια… Εξαρτάται και από το ν καθώς η κατανομή έχει διαφορετικό σχήμα για κάθε αριθμό δείγματος, όπως στο σχήμα παρακάτω





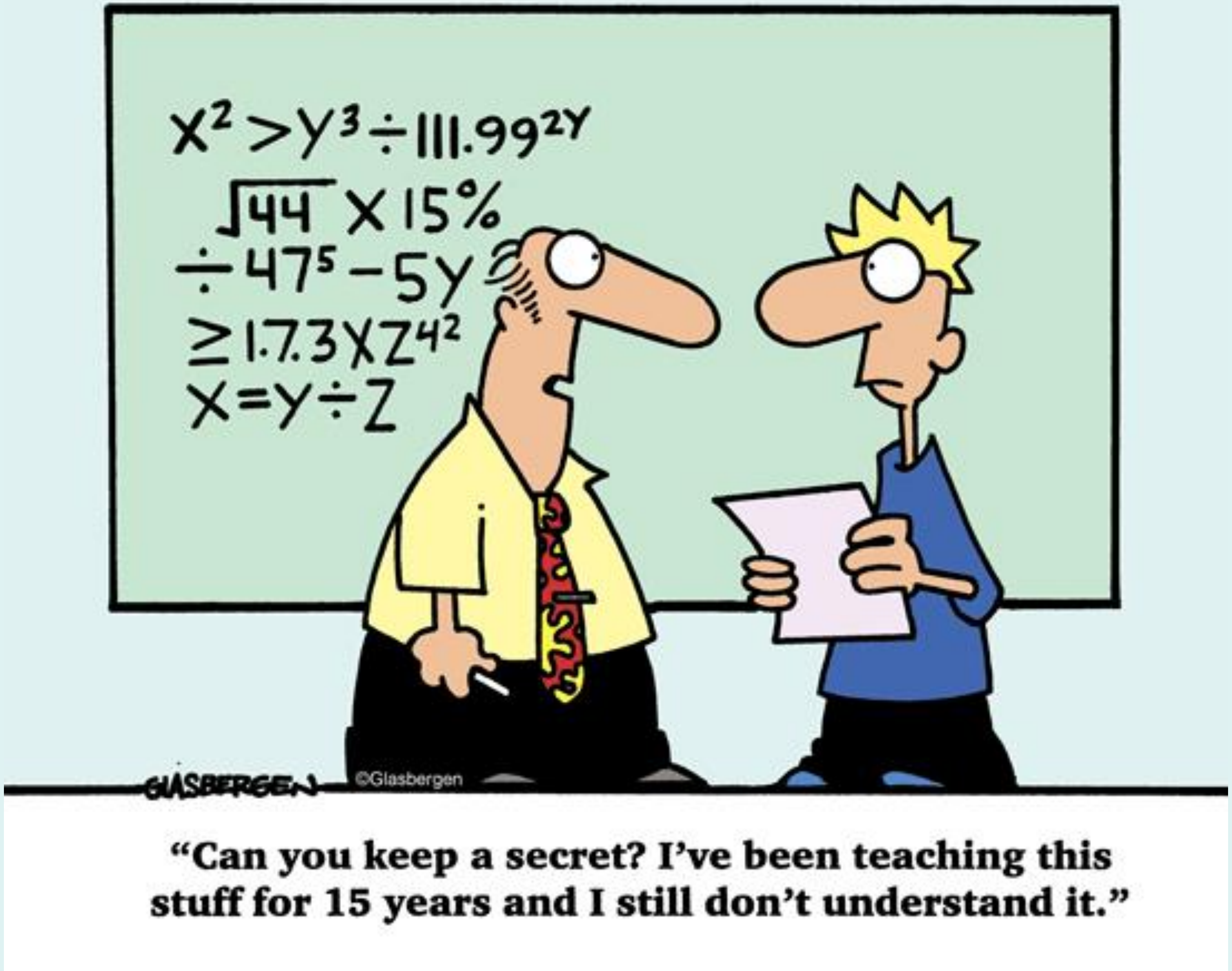


Κατανομή Gauss	-1.96 / +1.96	για επίπεδο εμπιστοσύνης 95%
Κατανομή Student για v=1	-12.706 /+12.706	για επίπεδο εμπιστοσύνης 95%
Κατανομή Student για n=2	-4.303 /+4.303	για επίπεδο εμπιστοσύνης 95%
Κατανομή Student για v=5	-2.78 /+2.78	για επίπεδο εμπιστοσύνης 95%
Κατανομή Student για v=20	-2.09 /+2.09	για επίπεδο εμπιστοσύνης 95%
Κατανομή Student για v=∞	-1.96 /+1.96	για επίπεδο εμπιστοσύνης 95% συμπίπτει με την κατανομή Gauss

Βλέπουμε πως το t μικραίνει όσο αυξάνονται και φτάνει προς το 1.96 της κανονικής κατανομής

Γιατί? Γιατί η κατανομή Gauss «ακουμπά» τον άξονα χ πιο νωρίς από όλες τις κατανομές Student, συνεπώς περιλαμβάνει γρηγορότερα την πιθανότητα π.χ. 95%. Η κατανομή Student είναι πιο πεπτλατυσμένη γιατί είναι πιο αβέβαιη

Ακόμη μια κατανομή που είναι χρήσιμη



Κατανομή F

Τι είναι η κατανομή F?

Είναι η κατανομή συχνότητας ενός λόγου $F = \frac{s_1^2}{s_2^2}$

Συνδέεται με την κατανομή x^2 (άλλη φορά)

Εξαρτάται από δυο βαθμούς ελευθερίας v_1 και v_2 και είναι μια συνεχής ασύμμετρη κατανομή.

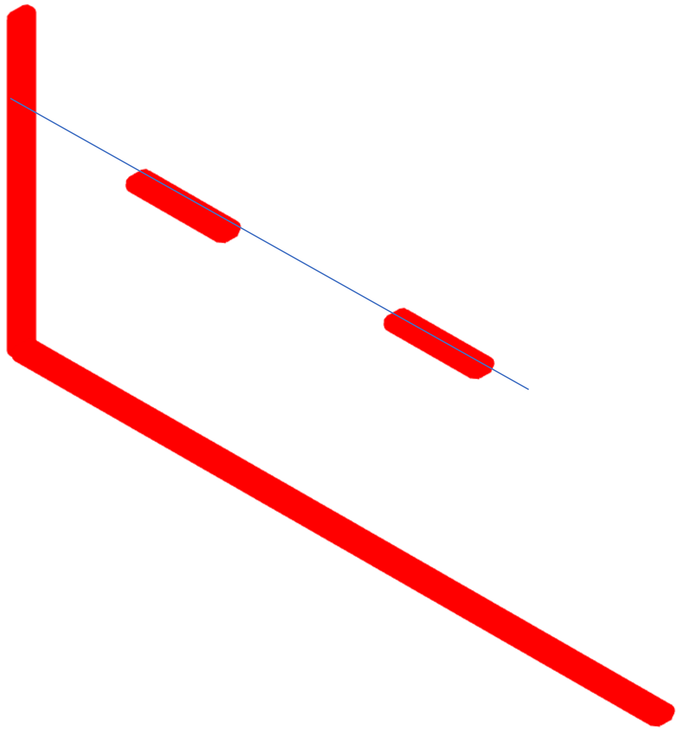
Το κυριότερο είναι ότι συνδέεται με το λόγο των αποκλίσεων δύο κατανομών

- Και τι μας νοιάζει αυτό άραγε?

Είναι πολύ σημαντικό για να συγκρίνουμε δύο κατανομές!!!!

Αν έχουν διαφορετικές αποκλίσεις τότε πρέπει να κάνουμε άλλες διαδικασίες απ’ ότι αν έχουν κοινή απόκλιση

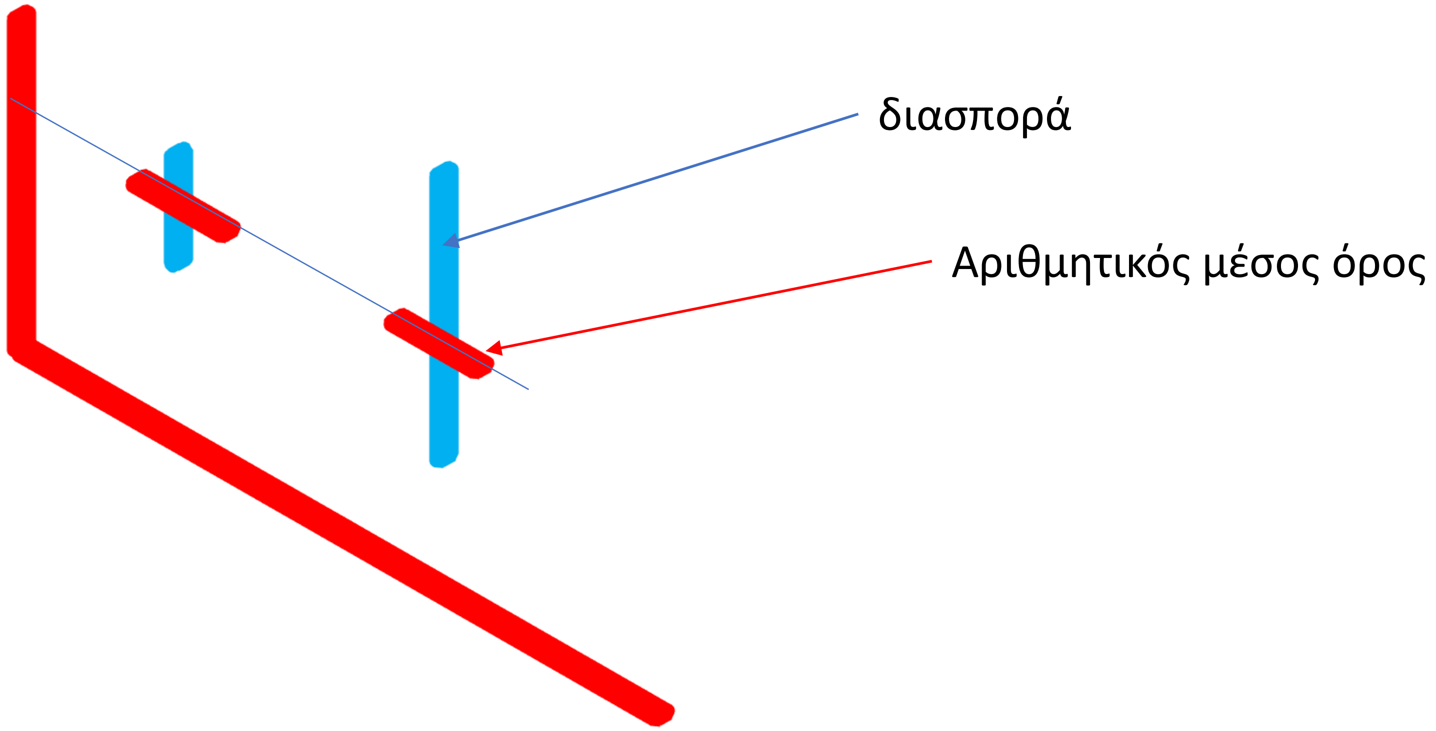
Ακόμα και αν δύο διαδικασίες (πειράματα) έχουν ίσο αριθμητικό μέσο όρο αυτό μπορεί να μην λέει και πολλά



Είναι πολύ σημαντικό για να συγκρίνουμε δύο κατανομές!!!!

Αν έχουν διαφορετικές αποκλίσεις τότε πρέπει να κάνουμε άλλες διαδικασίες απ’ ότι αν έχουν κοινή απόκλιση

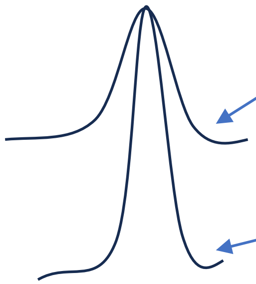
Ακόμα και αν δύο διαδικασίες (πειράματα) έχουν ίσο αριθμητικό μέσο όρο αυτό μπορεί να μην λέει και πολλά **αν έχουν διαφορετική διασπορά**



Είναι σημαντικό να συγκρίνουμε και τη διασπορά και γι αυτό θα χρησιμοποιήσουμε στη συνέχεια την κατανομή F

Γενική περίπτωση

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$



Κατανομή 1 με πληθυσμιακή απόκλιση σ_1 από την οποία παίρνουμε n1 δείγματα. Όταν υπολογίζουμε την δειγματική τυπική απόκλιση το αποτέλεσμα είναι s_1

Κατανομή 2 τα ίδια με την 1 πληθυσμιακή απόκλιση σ_2 n_2 δείγματα, δειγματική τυπική απόκλιση s_2

Ακόμα και αν έχουν τον ίδιο μέσο όρο, δεν μπορούμε να πούμε ότι οι δυο κατανομές είναι ίδιες αν δεν έχουν και ίδια τυπική απόκλιση. Τα λέω καλά?

Δηλαδή? Δηλαδή, στην περίπτωση της μεγαλύτερης τυπικής απόκλισης δεν έχουμε αρκετή ακρίβεια όσο στην άλλη κατανομή

Φυσικά οι δύο κατανομές θα μπορούσαν να έχουν διαφορετικούς μέσους όρους

Αυτή την απορία θα τη λύσει η κατανομή F

Στην παραπάνω εξίσωση συνήθως εξετάζουμε μια κατανομή δηλαδή $\sigma_1=\sigma_2$, π.χ. θέλουμε να δούμε αν από το ίδιο πείραμα ο αναλυτής 1 προσδιορίζει με την ίδια επαναληψιμότητα με τον αναλυτή 2. Η εξίσωση παίρνει μια πιο ειδική μορφή

$$F = \frac{s_1^2}{s_2^2}$$

Αυτή την μορφή συναντάμε συνήθως και θα ασχοληθούμε

Ποιες είναι οι παραδοχές για να έχω κατανομή F?

- Και οι δύο συγκρινόμενες κατανομές είναι κανονικές
- Και οι δύο πληθυσμοί είναι ανεξάρτητοι
- Λεπτομέρεια Η μεγαλύτερη απόκλιση ΠΑΝΤΑ βρίσκεται στον παρονομαστή – θα δούμε ότι η κατανομή είναι πάντα κυρτή προς την δεξιά πλευρά The larger sample variance always goes in the numerator to make the right-tailed test, and the right-tailed tests are always easy to calculate.

Η εξίσωση της κατανομής πιθανότητας της κατανομής F είναι

$$f_{n,m}(x) = \frac{\Gamma\left(\frac{n+m}{2}\right) n^{n/2}\, m^{m/2}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \; \frac{x^{n/2-1}}{(m+n\,x)^{(n+m)/2}}$$

Παρατηρούμε ότι εξαρτάται από τους βαθμούς ελευθερίας m και n

Να τονίσω πάλι ότι πρόκειται για σύγκριση δύο κατανομών με s_1^2 και s_2^2

Ο μέσος όρος είναι $\mu = \frac{m}{m-2}$

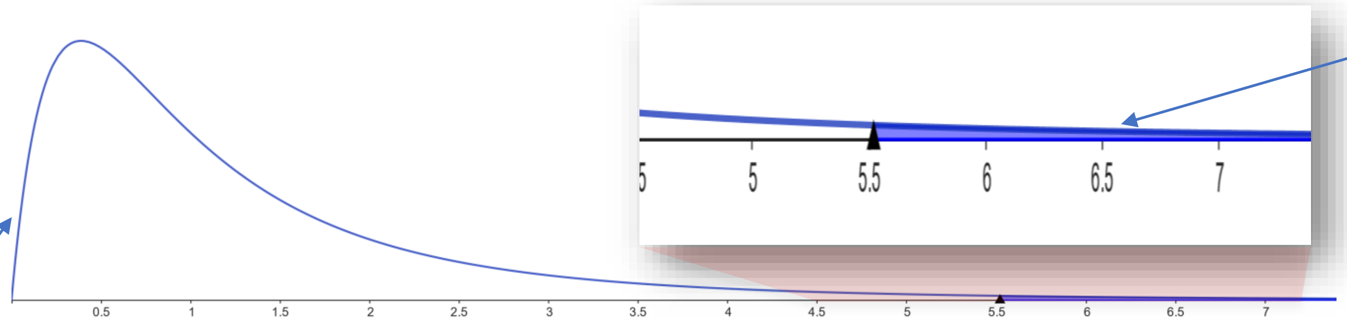
Η τυπική απόκλιση είναι $\sigma^2 = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$

Να τονίσω πάλι ότι πρόκειται για σύγκριση δύο δειγματοληψιών – κάθε μια έχει τους δικούς της βαθμούς ελευθερίας π.χ. στην πρώτη μπορεί κάποιος να έχει πάρει 5 δείγματα και στην δεύτερη 7. Οι βαθμοί ελευθερίας (ν-1) θα είναι 5-1=4 (δύσκολο μαθηματικό τερτίπι :P) για την πρώτη κατανομή και 7-1=6 για την δεύτερη.

Η κατανομή F έχει ένα περίεργο τρόπο για να βρίσκουμε το εμβαδό από την πυκνότητα πιθανότητας της κατανομής

Εμβαδό που δείχνει την πιθανότητα να συναντήσουμε τις αληθείς τιμές

Υπενθυμίζω με σχήμα για να μην φαίνεται κινέζικο



πυκνότητα πιθανότητας της κατανομής

F - Distribution (α = 0.01 in the Right Tail)

Ξεχωριστοί πίνακες για κάθε βαθμό σημαντικότητας εδώ 0.01

Numerator Degrees of Freedom

Δύο σειρές βαθμών ελευθερίας

$df_2 \backslash df_1$	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172

Denominator Degrees of Freedom

Μικρός αριθμός βαθμών ελευθερίας

Μεγάλος αριθμός βαθμών ελευθερίας

ΠΡΩΤΗ ΣΕΛΙΔΑ

F - Distribution (α = 0.01 in the Right Tail)

Ξεχωριστοί πίνακες για κάθε βαθμό σημαντικότητας εδώ 0.01

Numerator Degrees of Freedom (df_1); Denominator Degrees of Freedom (df_2)

df_2 \ df_1	10	12	15	20	24	30	40	60	120	∞
1	6055.8	6106.3	6157.3	6208.7	6234.6	6260.6	6286.8	6313.0	6339.4	6365.9
2	99.399	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	27.229	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	14.546	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	10.051	9.8883	9.7222	9.5526	9.4665	9.3793	9.2912	9.2020	9.1118	9.0204
6	7.8741	7.7183	7.5590	7.3958	7.3127	7.2285	7.1432	7.0567	6.9690	6.8800
7	6.6201	6.4691	6.3143	6.1554	6.0743	5.9920	5.9084	5.8236	5.7373	5.6495
8	5.8143	5.6667	5.5151	5.3591	5.2793	5.1981	5.1156	5.0316	4.9461	4.8588
9	5.2565	5.1114	4.9621	4.8080	4.7290	4.6486	4.5666	4.4831	4.3978	4.3105
10	4.8491	4.7059	4.5581	4.4054	4.3269	4.2469	4.1653	4.0819	3.9965	3.9090
11	4.5393	4.3974	4.2509	4.0990	4.0209	3.9411	3.8596	3.7761	3.6904	3.6024
12	4.2961	4.1553	4.0096	3.8584	3.7805	3.7008	3.6192	3.5355	3.4494	3.3608
13	4.1003	4.9603	3.8154	3.6646	3.5868	3.5070	3.4253	3.3413	3.2548	3.1654
14	3.9394	3.8001	3.6557	3.5052	3.4274	3.3476	3.2656	3.1813	3.0942	3.0040
15	3.8049	3.6662	3.5222	3.3719	3.2940	3.2141	3.1319	3.0471	2.9595	2.8684
16	3.6909	3.5527	3.4089	3.2587	3.1808	3.1007	3.0182	2.9330	2.8447	2.7528
17	3.5931	3.4552	3.3117	3.1615	3.0835	3.0032	2.9205	2.8348	2.7459	2.6530
18	3.5082	3.3706	3.2273	3.0771	2.9990	2.9185	2.8354	2.7493	2.6597	2.5660
19	3.4338	3.2965	3.1533	3.0031	2.9249	2.8442	2.7608	2.6742	2.5839	2.4893
20	3.3682	3.2311	3.0880	2.9377	2.8594	2.7785	2.6947	2.6077	2.5168	2.4212
21	3.3098	3.1730	3.0300	2.8796	2.8010	2.7200	2.6359	2.5484	2.4568	2.3603
22	3.2576	3.1209	2.9779	2.8274	2.7488	2.6675	2.5831	2.4951	2.4029	2.3055
23	3.2106	3.0740	2.9311	2.7805	2.7017	2.6202	2.5355	2.4471	2.3542	2.2558
24	3.1681	3.0316	2.8887	2.7380	2.6591	2.5773	2.4923	2.4035	2.3100	2.2107
25	3.1294	2.9931	2.8502	2.6993	2.6203	2.5383	2.4530	2.3637	2.2696	2.1694

Δύο σειρές βαθμών ελευθερίας

Μεγάλος αριθμός βαθμών ελευθερίας

Μικρός αριθμός βαθμών ελευθερίας

ΔΕΥΤΕΡΗ ΣΕΛΙΔΑ

F - Distribution (α = 0.025 in the Right Tail)

Numerator Degrees of Freedom

Denominator Degrees of Freedom

$df_2 \backslash df_1$	1	2	3	4	5	6	7	8	9
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
2	38.506	39.000	39.165	39.248	39.298	39.331	39.335	39.373	39.387
3	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473
4	12.218	10.649	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047
5	10.007	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6638	3.5879
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358
13	6.4143	4.9653	4.3472	3.9959	3.7667	3.6043	3.4827	3.3880	3.3120
14	6.2979	4.8567	4.2417	3.8919	3.6634	3.5014	3.3799	3.2853	3.2093
15	6.1995	4.7650	4.1528	3.8043	3.5764	3.4147	3.2934	3.1987	3.1227
16	6.1151	4.6867	4.0768	3.7294	3.5021	3.3406	3.2194	3.1248	3.0488
17	6.0420	4.6189	4.0112	3.6648	3.4379	3.2767	3.1556	3.0610	2.9849
18	5.9781	4.5597	3.9539	3.6083	3.3820	3.2209	3.0999	3.0053	2.9291
19	5.9216	4.5075	3.9034	3.5587	3.3327	3.1718	3.0509	2.9563	2.8801
20	5.8715	4.4613	3.8587	3.5147	3.2891	3.1283	3.0074	2.9128	2.8365
21	5.8266	4.4199	3.8188	3.4754	3.2501	3.0895	2.9686	2.8740	2.7977
22	5.7863	4.3828	3.7829	3.4401	3.2151	3.0546	2.9338	2.8392	2.7628
23	5.7498	4.3492	3.7505	3.4083	3.1835	3.0232	2.9023	2.8077	2.7313
24	5.7166	4.3187	3.7211	3.3794	3.1548	2.9946	2.8738	2.7791	2.7027
25	5.6864	4.2909	3.6943	3.3530	3.1287	2.9685	2.8478	2.7531	2.6766
26	5.6586	4.2655	3.6697	3.3289	3.1048	2.9447	2.8240	2.7293	2.6528

Ξεχωριστοί πίνακες για κάθε βαθμό σημαντικότητας

εδώ 0.025

Δύο σειρές βαθμών ελευθερίας

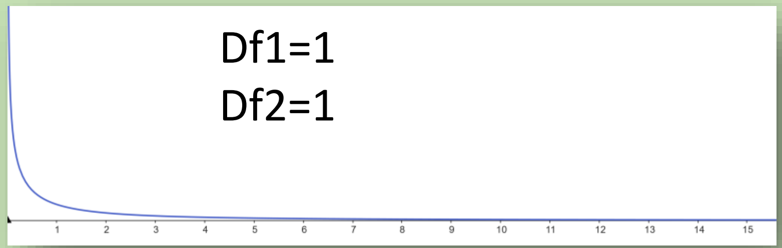
Μικρός αριθμός βαθμών ελευθερίας

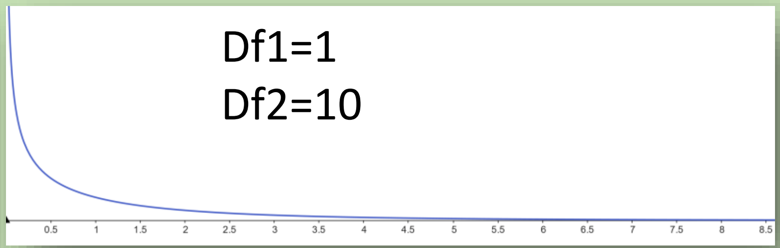
Μεγάλος αριθμός βαθμών ελευθερίας

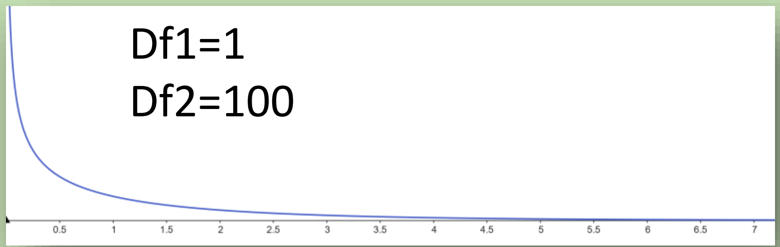
ΤΡΙΤΗ ΣΕΛΙΔΑ

(ουσιαστικά επανάληψη των δύο πρώτων σελίδων για άλλα επίπεδα σημαντικότητας)

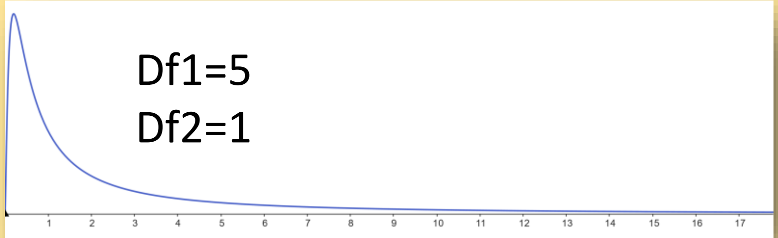
Η κατανομή F δεν έχει σταθερό σχήμα, αλλά το σχήμα εξαρτάται από του βαθμούς ελευθερίας των δύο κατανομών
Για να το δουμε
DF= degrees of freedom – βαθμοί ελευθερίας

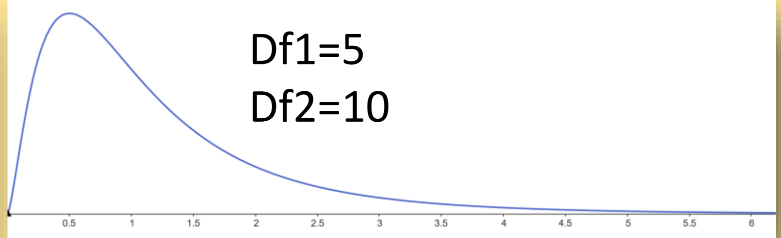


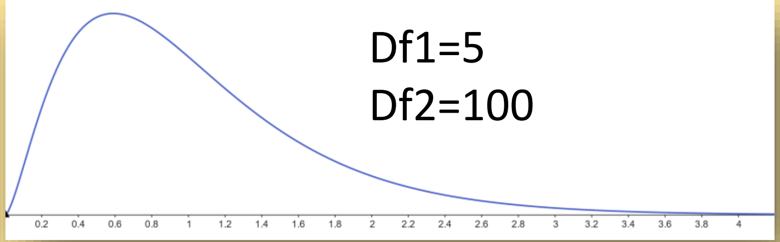




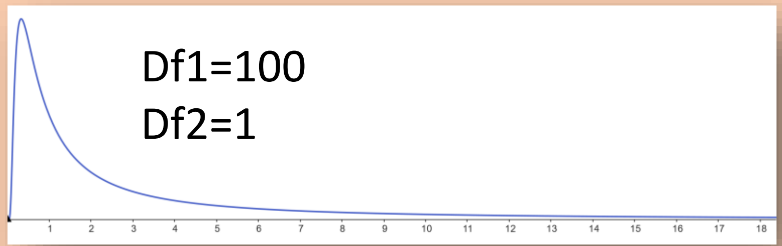
Η κατανομή δεν έχει κάποιο σημείο καμπής όπως οι άλλες που έχουμε δει
Οι βαθμοί ελευθερίας στον παρονομαστή δεν παίζουν πολύ σημαντικό ρόλο

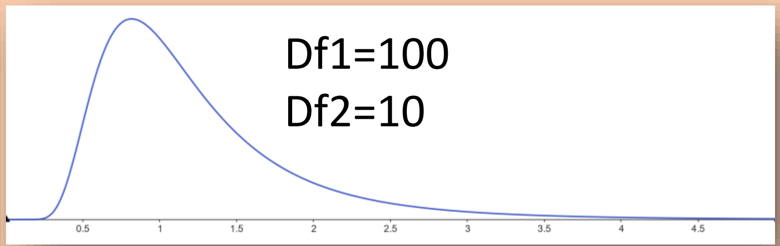


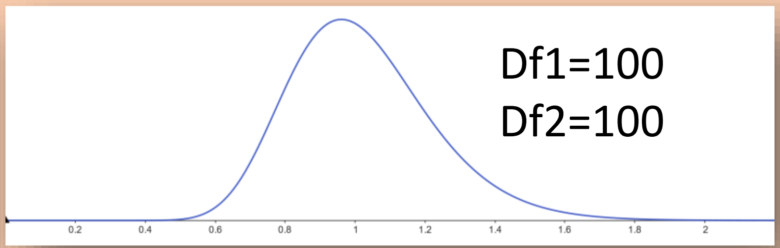




Η κατανομή έχει σημείο καμπής
Οι βαθμοί ελευθερίας στον αριθμητή καθορίζουν το σχήμα στον παρονομαστή το πλάτος







Το σημείο καμπής εξαρτάται από το Df1 και το πλάτος από το DF2.
Για μεγάλο αριθμό δειγμάτων η κατανομή F τείνει προς την κανονική κατανομή

Γιατί κατανομή F?

Προς τιμή του Fisher

Αλλά το όνομα έδωσε
ο George W. Snedecor