Ανάλυση δεδομένων στο περιβάλλον του SPSS

Λαβίδας Κωνσταντίνος

Μαθηματικός

lavidas@upatras.gr

1. Δειγματοληπτικές κατανομές

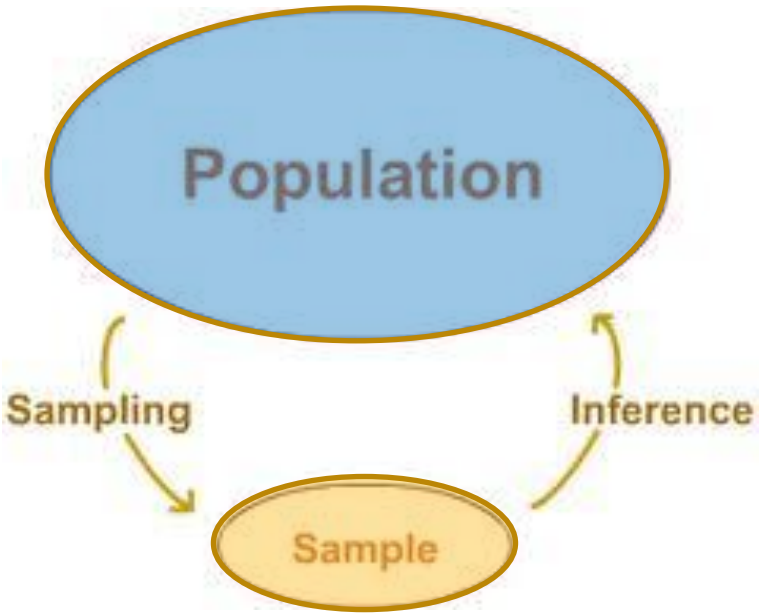
2. Σημειακή εκτίμηση και διαστήματα εμπιστοσύνης:

Εκτίμηση παραμέτρων του πληθυσμού

Προσδιορισμός μεγέθους δείγματος

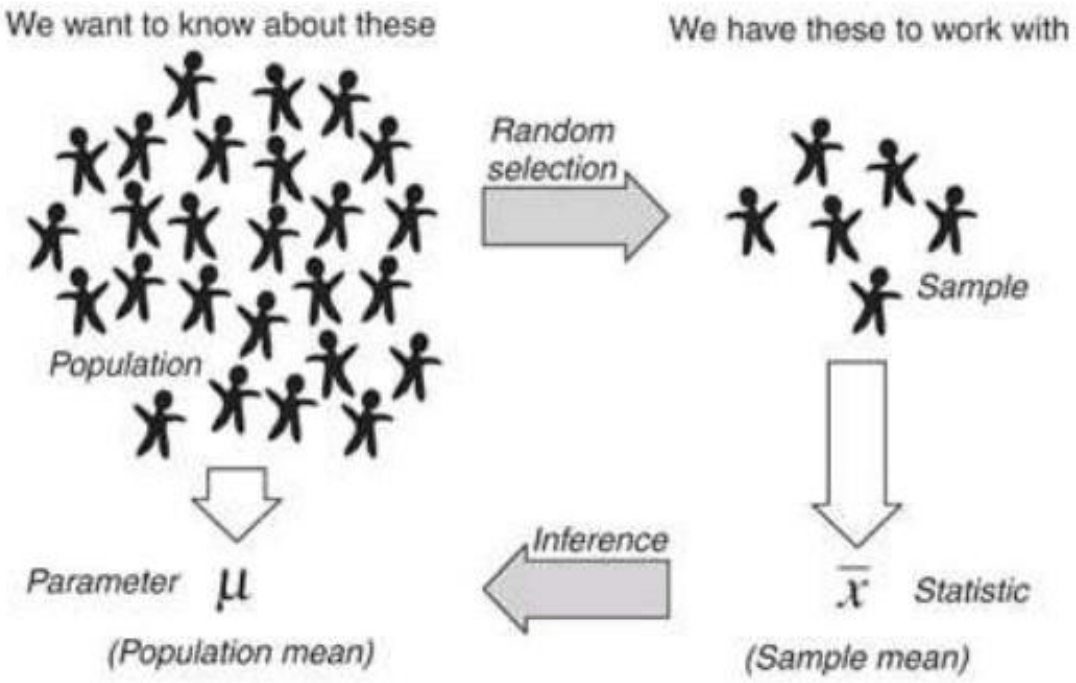
Επαγωγική Στατιστική Ανάλυση

- Είναι το στάδιο ανάλυσης των δεδομένων του δείγματος με το οποίο αποκτούμε πληροφορίες για να εξάγουμε συμπεράσματα για τον πληθυσμό (inference).



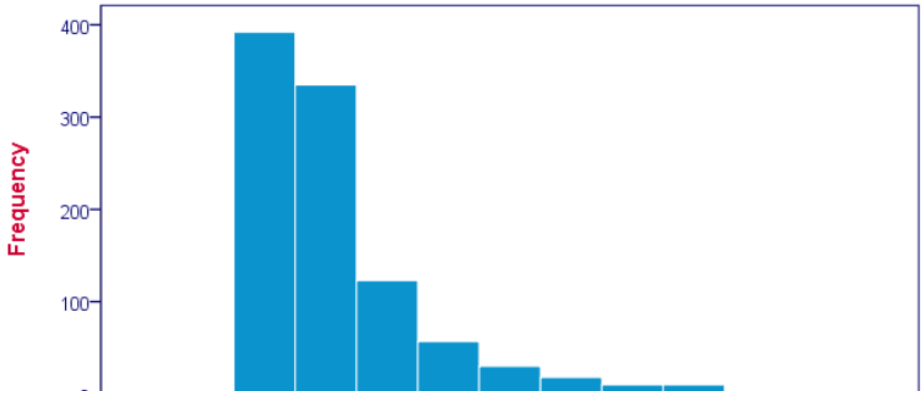
Παράδειγμα

- Έστω ότι μας ενδιαφέρει να προσδιορίσουμε την μέση επίδοση των πρωτοετών φοιτητών των τμημάτων ανθρωπιστικών σπουδών στην στατιστική.
- Προβληματισμοί:
  - Μάλλον δύσκολο να μετρήσω την μέση επίδοση σε όλο τον πληθυσμό. Γιατί;
  - Η σύνηθες τακτική να μετρήσω σε ένα υποσύνολο του πληθυσμού τη μέση τιμή της επίδοσης που βασίζεται στις δειγματικές μετρήσεις. Η μέση επίδοση του δείγματος πόσο θα απέχει από την αντίστοιχη του πληθυσμού;
  - Σε κάθε περίπτωση πως πρέπει να είναι το υποσύνολο αυτό; Τι μέγεθος πρέπει να έχει;


We want to know about these
We have these to work with
Random selection
Population
Sample
Parameter μ
(Population mean)
Inference
$\bar{x}$ Statistic
(Sample mean)

Παράδειγμα

- Ας υποθέσουμε ότι ο πληθυσμός της επίδοσης των 1220 πρωτοετών φοιτητών ανθρωπιστικών τμημάτων έχει μέση επίδοση 4,67 και έστω ότι η κατανομή τους παρουσιάζει μια έντονη θετική ασσυμετρία.



- Ας υποθέσουμε επίσης ότι δεν γνωρίζουμε την παραπάνω μέση τιμή της επίδοσης και επιθυμούμε να την βρούμε.
- Σκεφτόμαστε να επιλέξουμε ένα τυχαίο δείγμα, π.χ. 10 φοιτητών, να κάνουμε τις μετρήσεις μας και να βρούμε την μέση τιμή της επίδοσης σε αυτό το δείγμα.
- Έστω το δείγμα 10 φοιτητών: 2, 4, 6, 7, 8, 9, 10, 4, 5, 9 έχει ΜΤ=6,4 που είναι μεγαλύτερη από τη ΜΤ του πληθυσμού.

Παράδειγμα

- Η απόσταση που έχει η δειγματική μέση τιμή ή αλλιώς συνάρτηση (δείκτης) από την πληθυσμιακή συνάρτηση (παράμετρος) είναι το βασικό πρόβλημα στην επαγωγική στατιστική (**inferential statistics**).
- Η λύση είναι να προσδιορίσουμε λαμβάνοντας υπόψη τις δειγματικές μετρήσεις ένα πιθανό διάστημα τιμών (διάστημα εμπιστοσύνης) στο οποίο θα ανήκει η πληθυσμιακή συνάρτηση.

1. Δειγματοληπτική κατανομή

Δειγματοληπτική κατανομή

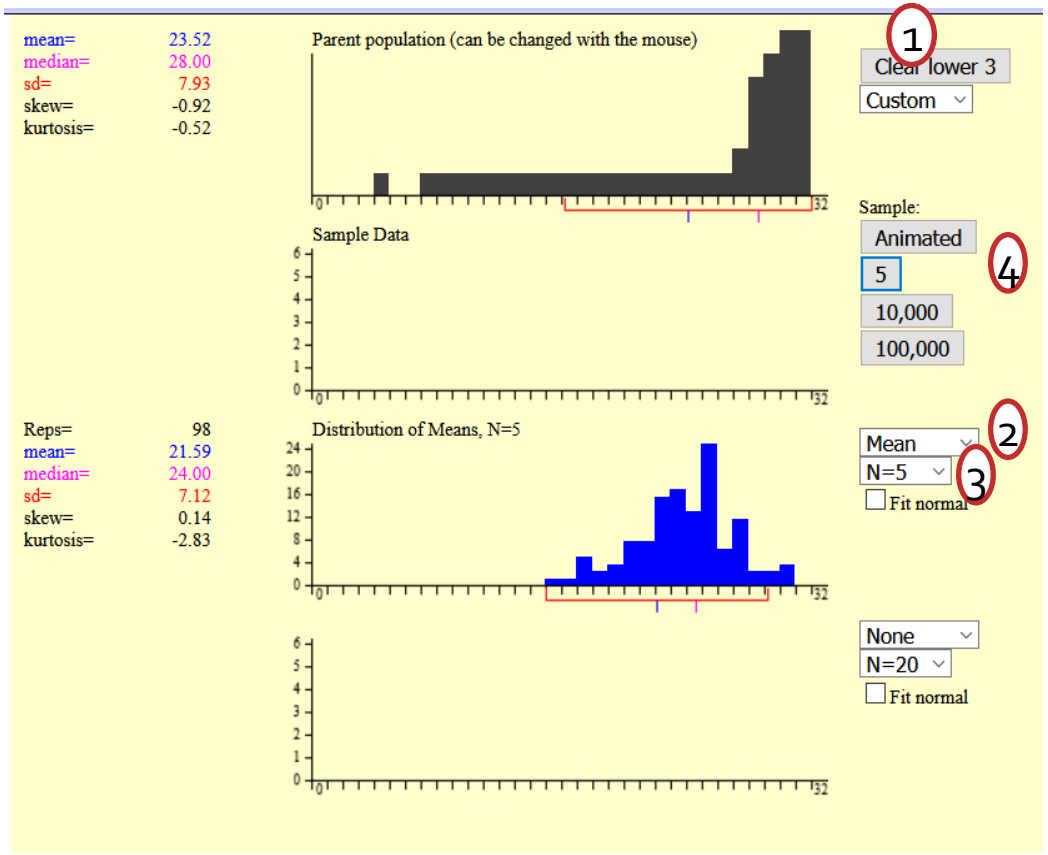
- Αν επιλέξουμε από έναν πληθυσμό π.χ. των 1220 πρωτοετών φοιτητών ανθρωπιστικών τμημάτων, πάρα πολλά δείγματα ίδιου μεγέθους (έστω 100).
- Και για κάθε ένα από αυτά να βρούμε την μέση τιμή του (επίδοση των πρωτοετών φοιτητών των τμημάτων ανθρωπιστικών σπουδών στην στατιστική):
  - Η κατανομή των μέσων τιμών όλων των παραπάνω δειγμάτων θα είναι λιγότερο ασύμμετρη από την πληθυσμιακή κατανομή.

Στον σύνδεσμο https://onlinestatbook.com/stat_sim/sampling_dist/

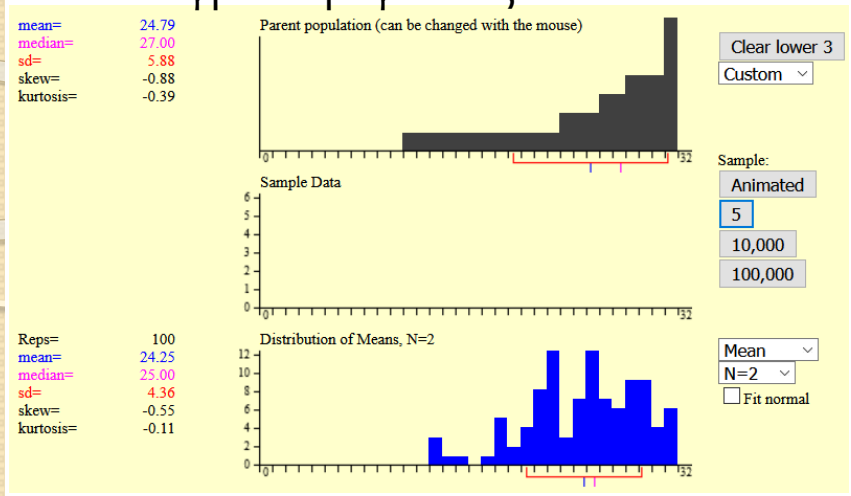
να δοκιμάσετε τα παραπάνω (προηγούμενη διαφάνεια): α) για δείγματα μεγέθους 2 και β) δείγματα μεγέθους 25, επιλέγοντας 100 δείγματα από τον υποτιθέμενο πληθυσμό.

- Τι παρατηρείτε;

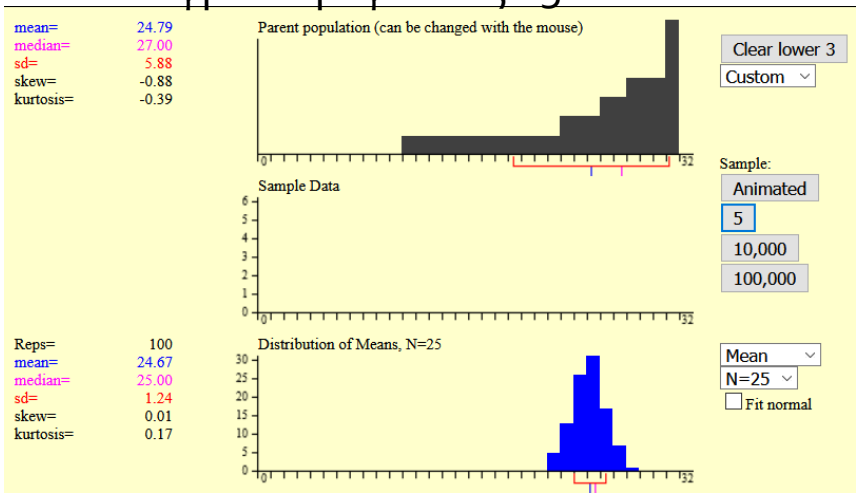
1. Αρχικά επιλέξτε ή σχεδιάστε την πληθυσμιακή κατανομή.
2. Επιλέξτε τη συνάρτηση (π.χ. Mean) η οποία θα υπολογίζεται για κάθε δείγμα τιμών από τον πληθυσμό.
3. Επιλέξτε το πλήθος των τιμών που θα έχουν τα τυχαία δείγματα (μέγεθος δείγματος) τα οποία θα επιλεγούν από τον παραπάνω πληθυσμό.
4. Επιλέξτε τον τρόπο που θα βλέπετε την προσομοίωση. Είτε ένα ένα τα στοιχεία του δείγματος (Animated) είτε 5 μαζί είτε ...



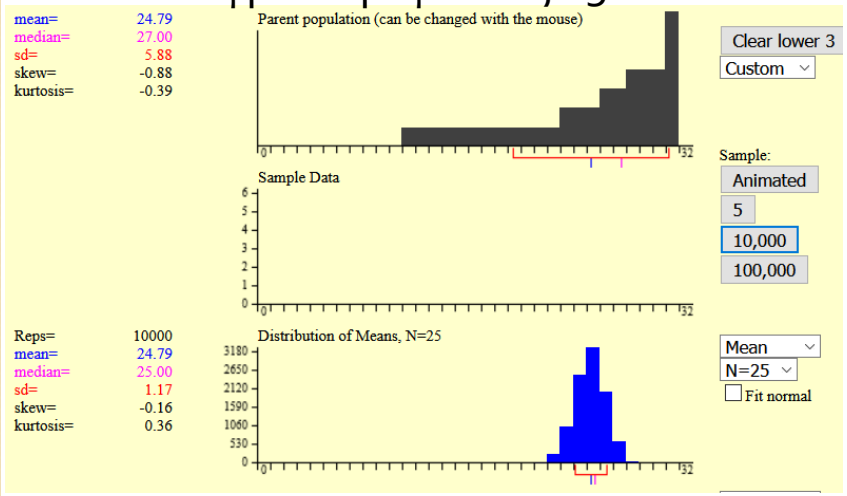
100 δείγματα μεγέθους 2



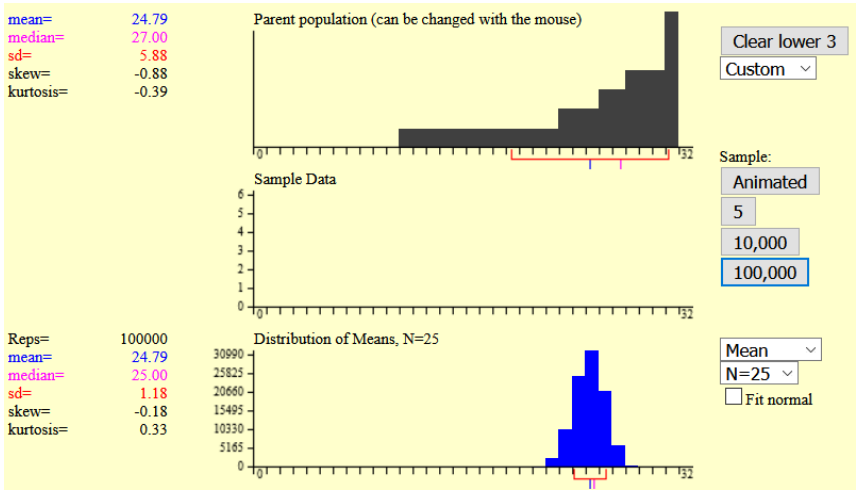
100 δείγματα μεγέθους 25



10000 δείγματα μεγέθους 25



100000 δείγματα μεγέθους 25



Απαραίτητα…

- Δειγματοληπτική κατανομή στατιστικών συναρτήσεων
- Κεντρικό Οριακό Θεώρημα (ΚΟΘ)

Δειγματοληπτική κατανομή στατιστικών συναρτήσεων

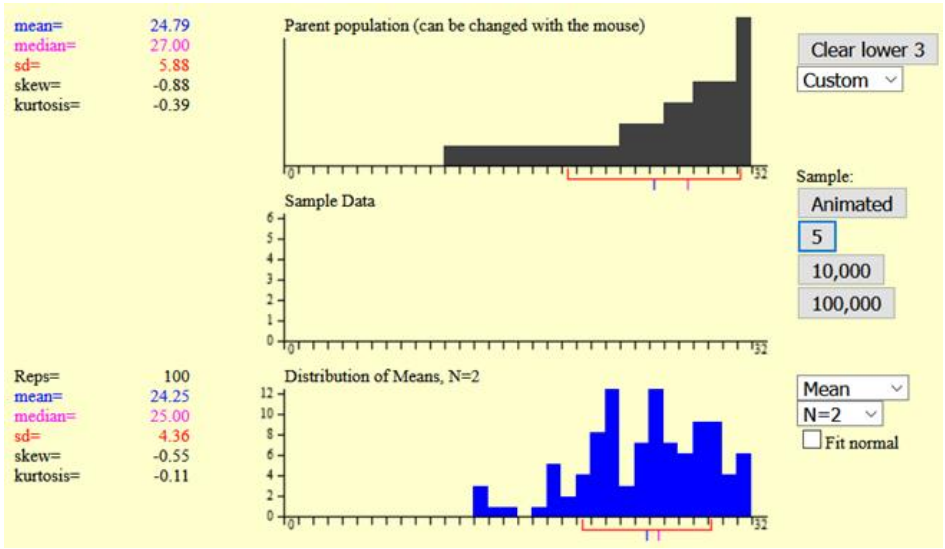
- Η δειγματοληπτική κατανομή (**sampling distribution**) μιας δειγματικής στατιστικής συνάρτησης (π.χ. μέση τιμή, διακύμανση, μια αναλογία, κ.λπ.) δημιουργείται αν θεωρητικά επιλέξουμε όλα τα δυνατά δείγματα ίδιου μεγέθους από ένα πληθυσμό και για κάθε ένα από αυτά υπολογίσουμε την στατιστική συνάρτηση.

- **Δειγματοληπτική κατανομή της δειγματικής μέσης τιμής ($\bar{X}$)**
- Αν επιλέξουμε όλα τα δυνατά δείγματα ίδιου μεγέθους (n) από έναν **κανονικά κατανεμημένο** πληθυσμό μεγέθους N, με **μέση τιμή (μ)** και **διακύμανση (σ^2),** τότε για τη δειγματοληπτική κατανομή της μέσης τιμής ($\bar{x}$) όλων των δυνατών δειγμάτων ισχύουν τα εξής:
  - είναι κανονικά κατανεμημένη
  - η μέση τιμή της δειγματοληπτικής κατανομής των μέσων τιμών ισούται με τη μέση τιμή (μ) του πληθυσμού
  - η διακύμανση της δειγματοληπτικής κατανομής ισούται με (σ^2/n).

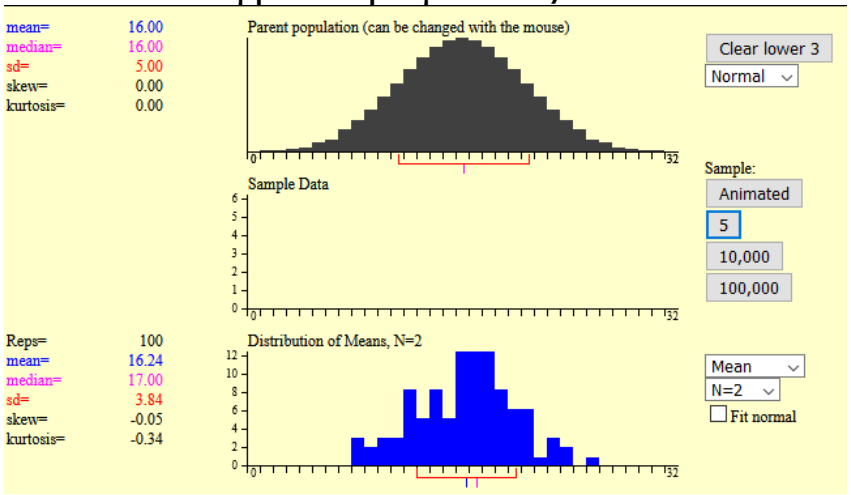
  Δηλαδή $\bar{X} \sim N(\mu, \sigma^2/n)$

- **Στατιστική συμπερασματολογία:**
  - Οι ιδιότητες της δειγματικής κατανομής της μέσης τιμής χρησιμοποιούνται στην εξαγωγή συμπερασμάτων που αφορούν τον πληθυσμό από τον οποίο προήλθε το δείγμα.

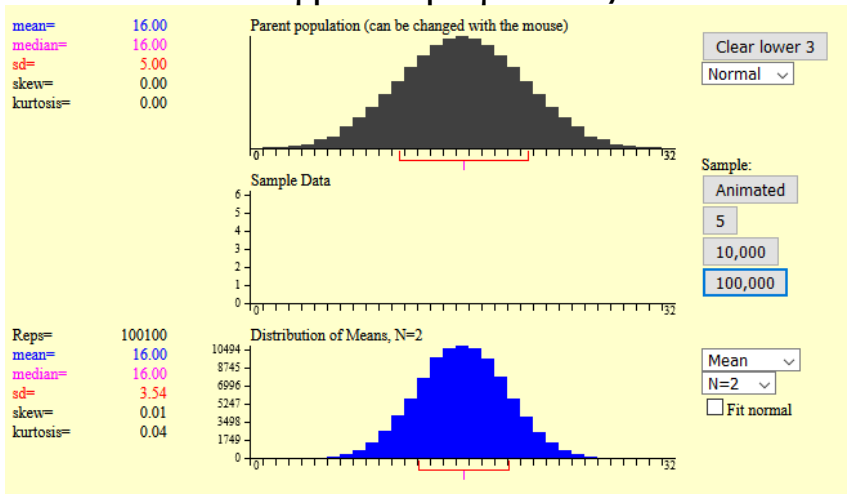
100 δείγματα μεγέθους 2



100 δείγματα μεγέθους 2



100100 δείγματα μεγέθους 2



Sd=5/1,41=3,54

Mean=16,00

Κεντρικό οριακό θεώρημα (ΚΟΘ) Central Limit Theorem (CLT)

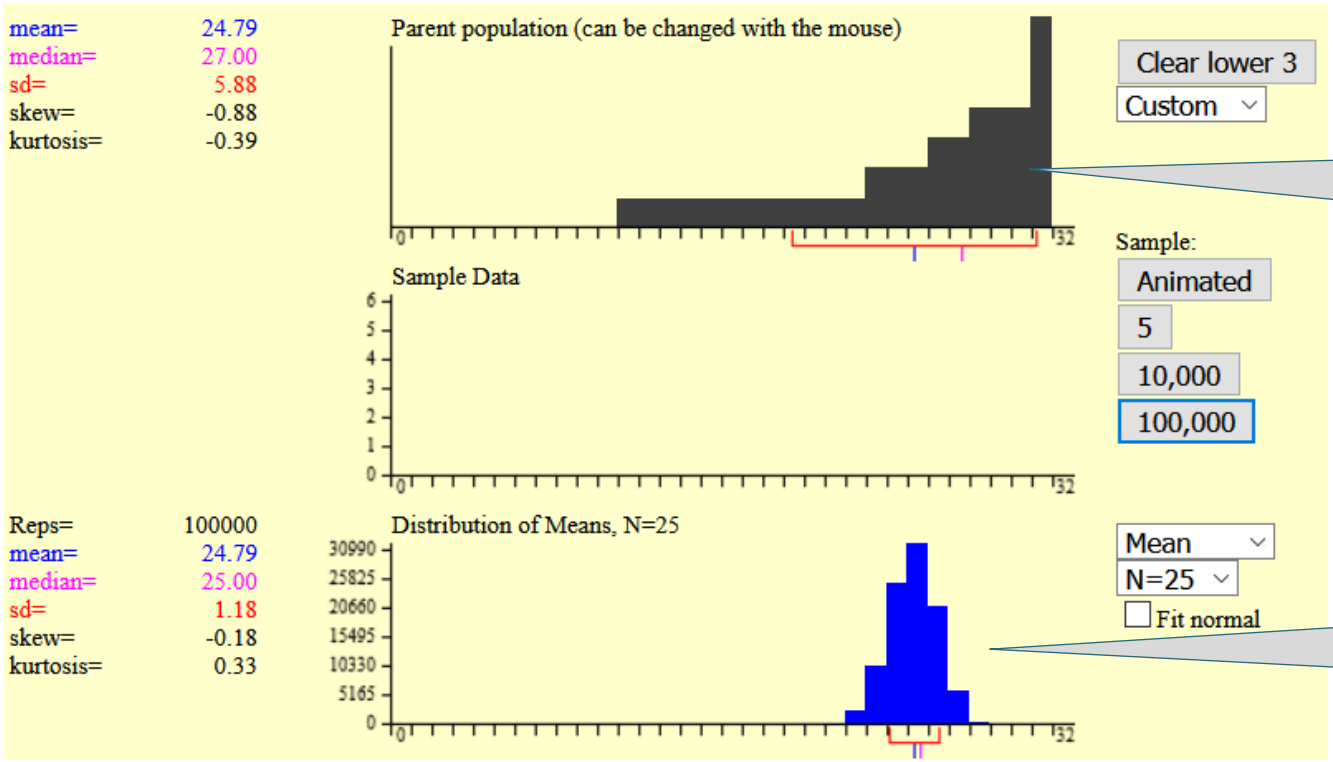
- Στην περίπτωση ενός πληθυσμού άγνωστης κατανομής (π.χ. μη κανονικά κατανεμημένου πληθυσμού):
- **ΚΟΘ:** Σε έναν οποιασδήποτε κατανομής πληθυσμό μεγέθους N, με μέση τιμή (μ) και τυπική απόκλιση (σ), η δειγματοληπτική κατανομή της δειγματικής μέσης τιμής όλων των δυνατών δειγμάτων μεγέθους (n), είναι:
  - κατά προσέγγιση κανονική
  - με μέση τιμή (μ) και διακύμανση σ^2/n),

  $\bar{X} \sim N(\mu, \sigma^2/n)$

  εφόσον το μέγεθος των δειγμάτων είναι επαρκώς μεγάλο:

  (Στις περισσότερες των περιπτώσεων $n \geq 30$).

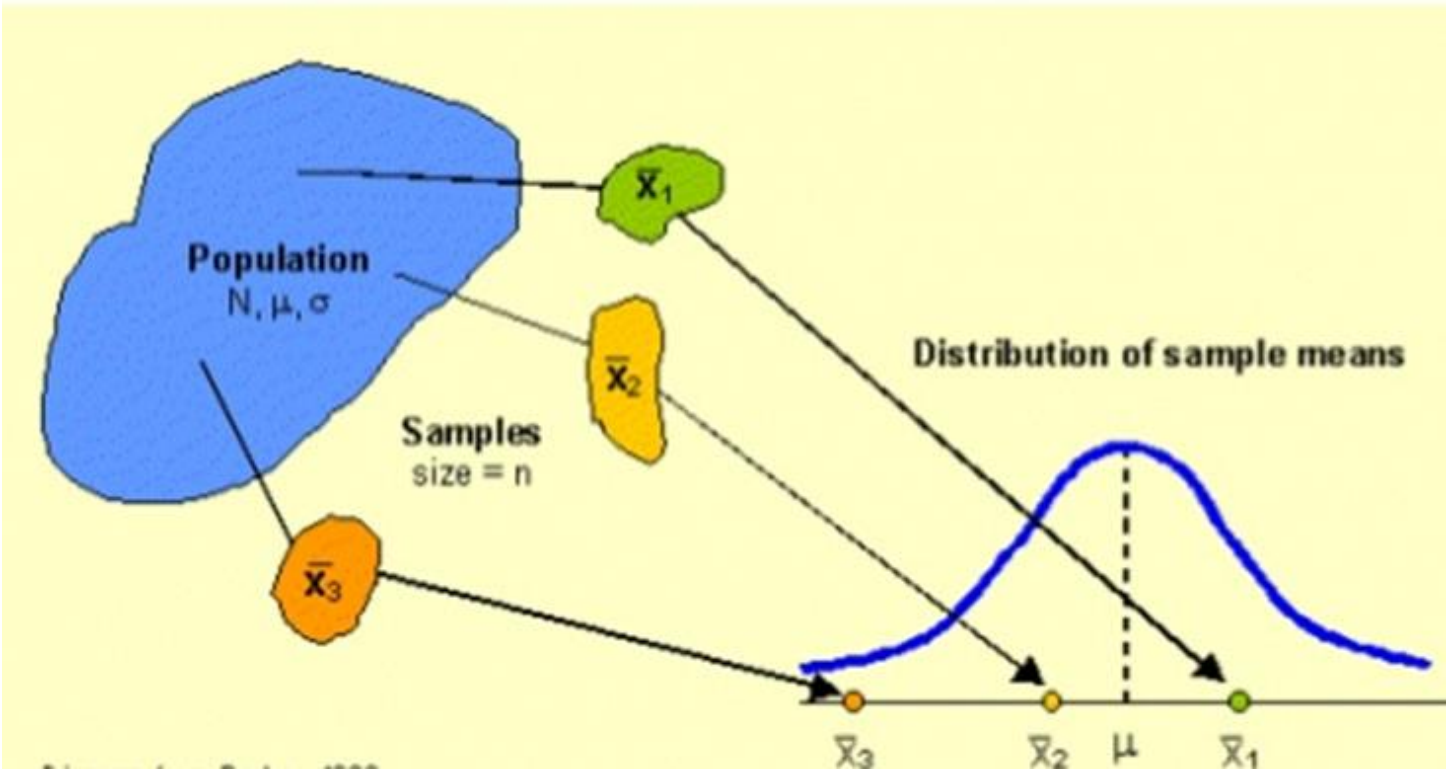
100000 δείγματα μεγέθους 25 (επαρκώς μεγάλα)



Μη κανονικά κατανεμημένος πληθυσμός

Κανονικά κατανεμημένη η δειγματοληπτκή κατανομή της μέσης τιμής

Θα πρέπει το μέγεθος του δείγματος να είναι επαρκώς μεγάλο (τουλάχιστον 30)



Δύο είδη στατιστικής συμπερασματολογίας

- Δύο κατευθύνσεις αξιοποίησης της επαγωγικής στατιστικής:
  - **η εκτιμητική** και
  - ο **έλεγχος υποθέσεων**.

2. Σημειακή εκτίμηση και διαστήματα εμπιστοσύνης:

- **Εκτίμηση παραμέτρων του πληθυσμού**
- **Προσδιορισμός μεγέθους δείγματος**

Εκτίμηση παραμέτρων του πληθυσμού

Εκτίμηση παραμέτρων του πληθυσμού

Προσδιορισμός **παραμέτρων** του πληθυσμού μέσω των **δεικτών** του δείγματος

- **Σημειακής εκτίμησης (point estimation)**
- **Εκτίμηση διαστήματος (interval estimation)**

Εκτίμηση παραμέτρων του πληθυσμού

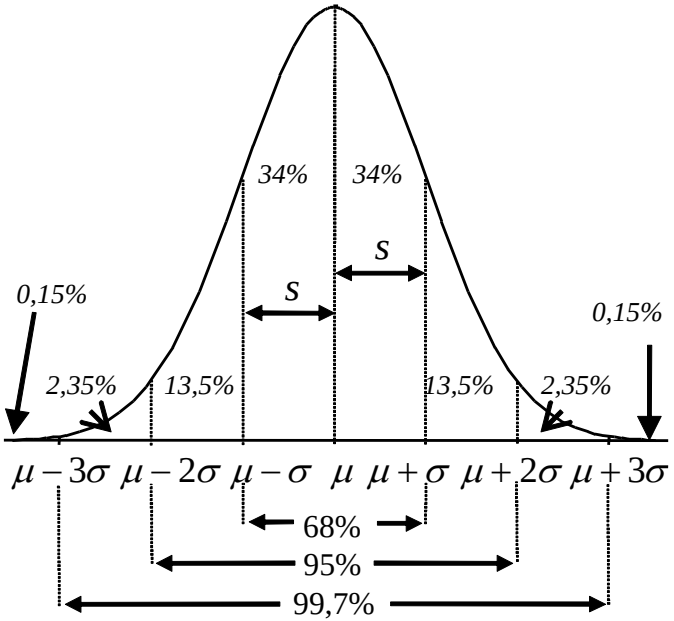
- **Σημειακής εκτίμησης (point estimation)**
  - Ο προσδιορισμός κάποιας παραμέτρου του πληθυσμού γίνεται μέσω ενός στατιστικού δείκτη.
    - Π.χ., η δειγματοληπτική μέση τιμή (δείκτης) χρησιμοποιείται για να ***εκτιμήσουμε*** την μέση τιμή του πληθυσμού (παράμετρος).
- Ο δειγματικός μέσος θεωρείται και **αμερόληπτος εκτιμητής** του πληθυσμιακού μέσου
- Προσοχή, η σημειακή εκτίμηση δεν προσδιορίζει ακριβώς την όποια παράμετρο αλλά κατά προσέγγιση.
  - Μάλιστα μαζί με τον εκτιμητή πρέπει να παρουσιάζεται και η τυπική απόκλιση αλλά και το μέγεθος του δείγματος.

Εκτίμηση παραμέτρων του πληθυσμού

- **Εκτίμηση διαστήματος (interval estimation)**
  - Περιέχει ένα **διάστημα δυνατών τιμών** μέσα στο οποίο περιλαμβάνονται πιθανόν οι τιμές της παραμέτρου του πληθυσμού
  - **Διάστημα εμπιστοσύνης (ΔΕ)** (confidence interval):
    - σε επίπεδο εμπιστοσύνης ή ακρίβειας: **1-α**.
    - Όπου α=το επίπεδο σημαντικότητας ή η πιθανότητα να κάνουμε λάθος και το ΔΕ να μην περιέχει την εκτιμώμενη τιμή της παραμέτρου του πληθυσμού.
      - Συνήθως η πιθανότητα αυτή είναι α=5% άρα το επίπεδο εμπιστοσύνης είναι 95%.
  - Το σύνηθες ΔΕ που υπολογίζουμε είναι το **95% διάστημα εμπιστοσύνης**

Παράδειγμα

Σε μια δειγματοληπτική κατανομή, ένα τυχαίο δείγμα κ, από τα πολλά που είχατε επιλέξει (όλα μεγέθους 25) έχει μέση τιμή $\bar{x}_k = 4$. Αν η μέση τιμή της δειγματοληπτικής κατανομής είναι μ και η τυπική της απόκλιση είναι σ= $\frac{3}{\sqrt{25}} = 0{,}6$, να υπολογίσετε ένα διάστημα δυνατών τιμών της δειγματοληπτικής κατανομής, μέσα στο οποίο περιλαμβάνονται το 95% των τιμών της κατανομής, συμμετρικά τοποθετημένες γύρω από τη μέση τιμή $\bar{x}_k$=4 του παραπάνω τυχαίου δείγματος.



Αν το δείγμα αυτό έχει μέση τιμή που συμπίπτει με τη μέση τιμή μ του πληθυσμού , ποιο θα είναι το ποσοστό των δειγμάτων που οι μέσες τιμές τους $\bar{x}_i$ απέχουν από το $\bar{x}_\kappa$ έως και περίπου δύο τυπικές αποκλίσεις ή P(|$\bar{x}_i - \bar{x}_\kappa$ |<=2σ); Η απάντηση είναι περίπου 95%.

Επομένως για το 95% των δειγμάτων ισχύει περίπου

$$-2\sigma \le \bar{x}_i - \bar{x}_\kappa \le 2\sigma \quad \text{ή}$$

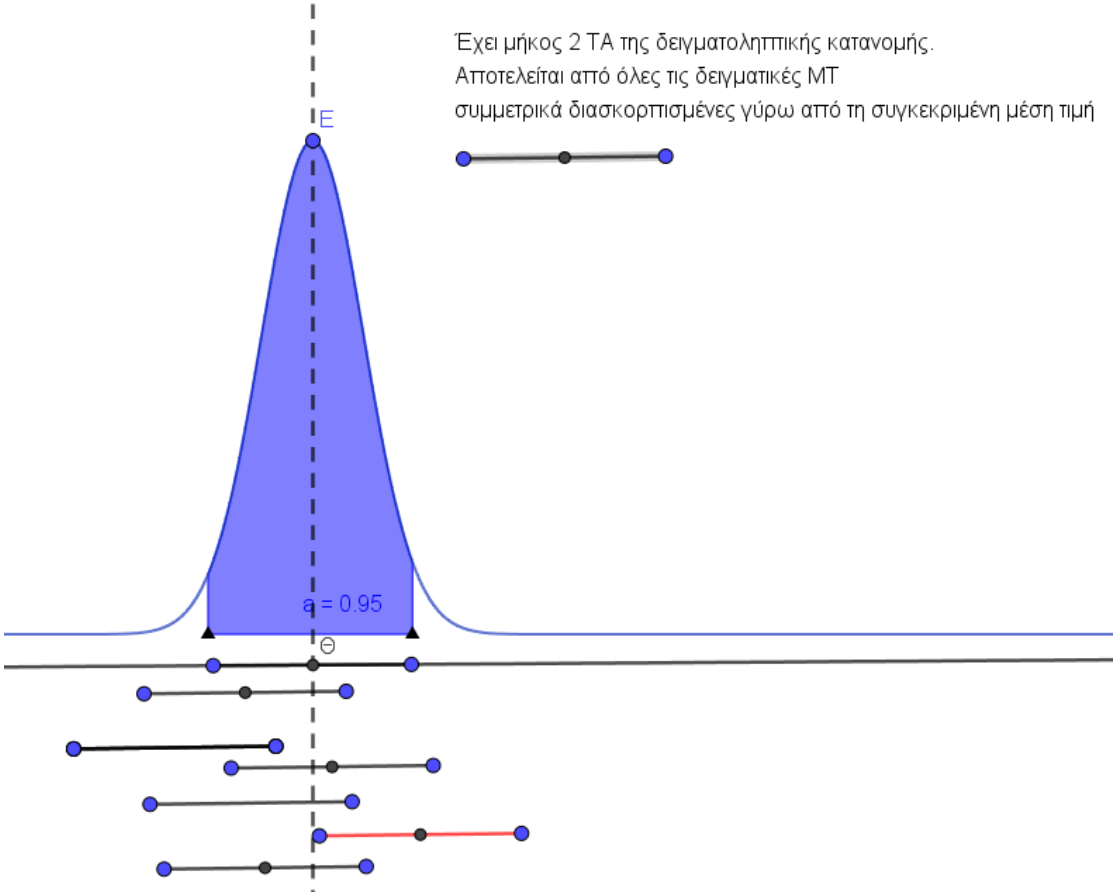
$$\bar{x}_\kappa - 2\sigma \le \bar{x}_i \le \bar{x}_\kappa + 2\sigma$$

Το τελευταίο είναι διάστημα των δυνατών τιμών της κατανομής στο οποίο περιλαμβάνονται το 95% των δειγματικών μέσων της κατανομής, συμμετρικά τοποθετημένες γύρω από τη μέση τιμή $\bar{x}_\kappa$ του παραπάνω τυχαίου δείγματος

Συνεπώς <u>για το δικό μας δείγμα</u> το διάστημα αυτό είναι

$$\bar{x}_\kappa - 2\sigma \le \bar{x}_i \le 2\sigma + \bar{x}_\kappa \ \acute{\eta}$$

$$[\bar{x}_\kappa - 2\sigma,\ \ 2\sigma + \bar{x}_\kappa] \ \acute{\eta}\ 2{,}8 \le \bar{x}_i \le 4{,}6$$



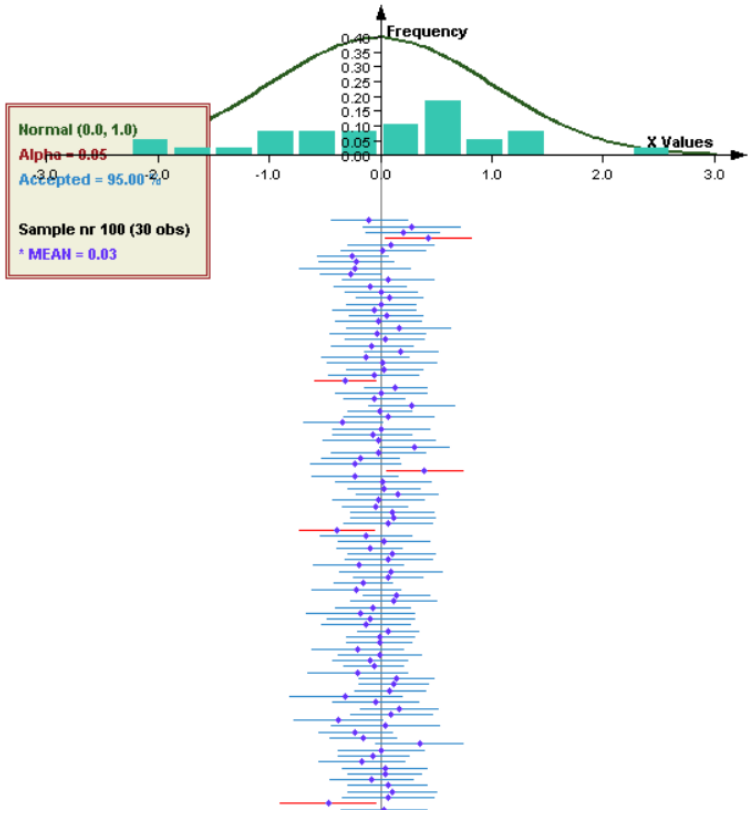
Κάθε φορά φτιάχνεις ένα διάστημα δειγματικών μέσων που είναι συμμετρικά διασκορπισμένες γύρω από την συγκεκριμένη δειγματική μέση τιμή (τελεία στο μέσο του ευθυγράμμου τμήματος).

Το διάστημα αυτό έχει μήκος 4 ΤΑ. Όσο δηλαδή και το διάστημα που αντιστοιχεί στο 95% των δειγματικών μέσων με κέντρο τη ΜΤ του πληθυσμού.

https://www.geogebra.org/t/confidence-interval

95% διάστημα εμπιστοσύνης της μέσης τιμής του πληθυσμού

- Αν πάρουμε 100 δείγματα ίδιου μεγέθους από τον ίδιο πληθυσμό, τότε περιμένουμε ότι στα 95 από αυτά τα αντίστοιχα διαστήματα εμπιστοσύνης να περιέχουν τη μέση τιμή του πληθυσμού.
- Και 5 να μην τον περιέχουν



Εκτιμητική με Εκτίμηση Διαστήματος

- Για παράδειγμα, υποθέστε ότι θέλουμε να εκτιμήσουμε τη μέση διδακτική εμπειρία των αδιόριστων εκπαιδευτικών, ειδικότητας φυσικών επιστημών.
- Χρησιμοποιούμε ένα δείγμα εκπαιδευτικών (*n*=30), οι οποίοι έχουν μέση διδακτική εμπειρία 11 μήνες.
- Αφού το μέγεθος του δείγματος είναι τουλάχιστον 30 σύμφωνα με το ΚΟΘ η δειγματοληπτική κατανομή της μέσης τιμής της διδακτικής εμπειρίας θα είναι κατά προσέγγιση κανονική
- Επομένως μέσω της μέσης τιμής ($\bar{x}$) του δείγματος: 11 μήνες
  - Αν γνωρίζουμε και την τυπική απόκλιση του πληθυσμού (σ)
  - Μέσω της τυπικής κανονικής κατανομής
- Θα προσδιοριστεί ένα διάστημα (**διαστημική εκτίμηση)** της μέσης τιμής του πληθυσμού
- Προσδιορισμός του **διαστήματος εμπιστοσύνης**:
  - Το διάστημα στο οποίο θα βρίσκεται με βεβαιότητα **95%** (η πιθανότητα να βρίσκετε στο διάστημα αυτό είναι 95%) η μέση τιμή (μ) του πληθυσμού

- Μέσης τιμής του δείγματος ($\bar{x}$)
- Τυπική απόκλιση του πληθυσμού (σ)
- Μέση τιμή του πληθυσμού (μ)

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \left(\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Ζητούμενη ποσότητα

Προσδιορισμός μεγέθους δείγματος

Προσδιορισμός του Μεγέθους του Δείγματος για εκτίμηση της παραμέτρου – Γρήγορος υπολογισμός

Ο Τύπος του Slovin

$n=N/(1+Ne^2)$.

n = μέγεθος δείγματος, N = μέγεθος του πληθυσμού e = λάθος τύπου α.

Αν επιθυμείς να προσδιορίσεις την παράμετρο ενός πληθυσμού με ένα διάστημα εμπιστοσύνης 95% τότε α=1-95%.

https://www.statisticshowto.com/how-to-use-slovins-formula/

Προσδιορισμός του Μεγέθους του Δείγματος για εκτίμηση της παραμέτρου: μέσης τιμής

- Υποθέστε ότι θέλουμε να εκτιμήσουμε την **μέση τιμή (μ) του πληθυσμού** μεγέθους N για τον οποίο γνωρίζουμε την τυπική απόκλιση (σ) αλλά και το δειγματοληπτικό σφάλμα d.
- Το ερώτημα που απασχολεί τον ερευνητή είναι: **ποιο είναι το μέγεθος του δείγματος (n=;)** που θα επιλέξουμε έτσι ώστε να έχουμε την καλύτερη προσέγγιση με το ελάχιστο κόστος.
- $\boldsymbol{d = |\mu - \bar{x}|}$ είναι το δειγματοληπτικό σφάλμα προσδιορισμού της παραμέτρου.
  - Αφού $(\bar{x} - z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}})$ αντιστοιχεί στο ποσοστό (1-α)100 % της κατανομής στο οποίο βρίσκεται **η μέση τιμή του πληθυσμού τότε για την μέγιστη τιμή του σφάλματος** $\boldsymbol{d = z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}}$**, λύνουμε ως προς n.**
  - Έτσι η αρχική προσέγγιση του μεγέθους n δείγματος είναι $n_0 = (\frac{z_{a/2}\sigma}{d})^2$
  - Τελικώς το μέγεθος του δείγματος προκύπτει :
  - 1η περίπτωση αν $\frac{n_0}{N} \leq 0{,}05$ (το πολύ 5% του πληθυσμού) τότε $n = n_0$
  - 2η περίπτωση αν $\frac{n_0}{N} > 0{,}05$ τότε $n = \frac{n_0}{1+\frac{n_0}{N}}$
  - Προσοχή:
    - Αν το μέγεθος **N** του πληθυσμού είναι θεωρητικά άπειρο (συνήθως για μεγαλύτερο του 10.000) τότε χρησιμοποιούμε την 1η περίπτωση.
    - Αν η τυπική απόκλιση του πληθυσμού δεν είναι γνωστή:
      - Τότε σε τυχαίο δείγμα μεγέθους μεγαλύτερο από 30 βρίσκουμε την τυπική απόκλιση του δείγματος και την χρησιμοποιούμε στη θέση της τυπικής απόκλισης του πληθυσμού.
      - Εναλλακτικά θεωρώντας ότι οι τιμές της μεταβλητής που αναζητούμε την μέση τιμή στον πληθυσμό κατανέμονται κανονικά, ισχύει ότι $\frac{max-min}{6} \cong \sigma$.

Προσδιορισμός του Μεγέθους του Δείγματος για εκτίμηση της παραμέτρου: μέσης τιμής (2)

- Παράδειγμα
  - Θέλουμε να εκτιμήσουμε τη μέση επίδοση μαθητών σε ένα τεστ αναγνωστικής ετοιμότητας. Λαμβάνοντας υπόψη ότι το δειγματοληπτικό σφάλμα του προσδιορισμού της παραμέτρου θέλουμε να είναι ±0, 5, η τυπική απόκλιση του πληθυσμού η οποία προσδιορίστηκε δειγματοληπτικά είναι περίπου s=2 και το μέγεθος του πληθυσμού είναι N=2000, να υπολογίσετε το μέγεθος του δείγματος που χρειάζεται χρησιμοποιώντας ένα 99% διάστημα εμπιστοσύνης.
- $\boldsymbol{d = z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}}}$: d=0, 5, σ=2, α=1-99%, $\frac{\alpha}{2} = 0{,}005$ και επομένως από τον πίνακα τυπικής κανονικής κατανομής: **z=2,575**
- Αρχικά $n_0 = (\frac{z_{a/2} \cdot \sigma}{d})^2 \Leftrightarrow n_0 = (\frac{2{,}58 \cdot 2}{0{,}5})^2 \cong 106$
- Αφού $\frac{n_0}{N} = \frac{106}{2000} = 0{,}053 > 0{,}05$ τότε
- $n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{106}{1{,}53} \cong 101$ μέγεθος δείγματος για την παραπάνω διερεύνηση

Προσδιορισμός του Μεγέθους του Δείγματος για εκτίμηση της παραμέτρου: ποσοστό (1)

- Αν θέλουμε να προσδιορίσουμε το ποσοστό (p) του πληθυσμού μεγέθους N που ανήκει σε κάποια από τις κ κατηγορίες μιας μεταβλητής, γνωρίζοντας το δειγματοληπτικό σφάλμα d.
- Το ερώτημα που απασχολεί τον ερευνητή είναι: **ποιο είναι το μέγεθος του δείγματος** που θα επιλέξουμε έτσι ώστε να έχουμε την καλύτερη προσέγγιση με το ελάχιστο κόστος.
- $\boldsymbol{d = |p - \hat{p}|}$ είναι το δειγματοληπτικό σφάλμα προσδιορισμού της παραμέτρου.
  - Αφού $(\boldsymbol{\hat{p} - z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}}, \hat{p} + z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}}})$ είναι το διάστημα (1-α)100% στο οποίο βρίσκεται το ποσοστό **του πληθυσμού τότε** $\boldsymbol{d = z_{\frac{a}{2}} \sqrt{\frac{(1-\hat{p})\hat{p}}{n}}}$**, λύνουμε ως προς n.**
  - Έτσι η πρώτη προσέγγιση του μεγέθους n δείγματος είναι $n_0 = (\frac{z_{\frac{a}{2}}}{d})^2 (1-\hat{p})\hat{p}$

Τελικώς το μέγεθος του δείγματος προκύπτει :

- 1η περίπτωση αν $\frac{n_0}{N} \leq 0{,}05$ (το πολύ 5% του πληθυσμού) τότε $n = n_0$
- 2η περίπτωση αν $\frac{n_0}{N} > 0{,}05$ τότε $n = \frac{n_0}{1+\frac{n_0-1}{N}}$
- Όπου **N** το μέγεθος του Πληθυσμού αναζήτησης της παραμέτρου.
- Προσοχή:
- Αν το μέγεθος **N** του δείγματος είναι θεωρητικά άπειρο (μεγαλύτερο του 10.000) τότε χρησιμοποιούμε την 1η περίπτωση.
- Αν το δειγματοληπτικό ποσοστό δεν είναι γνωστό, τότε θεωρούμε ως ποσοστό την τιμή του $\boldsymbol{\hat{p} = 1/2}$ για την οποία η ποσότητα $(1-\hat{p})\hat{p}$ γίνεται μέγιστη.

Προσδιορισμός του Μεγέθους του Δείγματος για εκτίμηση της παραμέτρου: ποσοστό (2)

- Έστω ότι θέλουμε να προσδιορίσουμε το ποσοστό (p) του μαθητικού πληθυσμού μεγέθους N=750 που αξιολογείται «θετικά» από το ΚΕΔΔΥ. Λαμβάνοντας υπόψη ότι το δειγματοληπτικό σφάλμα του προσδιορισμού της παραμέτρου θέλουμε να είναι ±0,05 (5%), να υπολογίσετε το μέγεθος του δείγματος που χρειάζεται χρησιμοποιώντας ένα 95% διάστημα εμπιστοσύνης.
- $\boldsymbol{d = z_{\frac{a}{2}}\sqrt{\frac{(1-\hat{p})\hat{p}}{n}}}$, d=0,05, α=1-95%=0,05, $\frac{\alpha}{2} = 0{,}025$ και επομένως από τον πίνακα τυπικής κανονικής κατανομής: z=1,96
- Αρχικά $n_0 = (\frac{z_{\frac{a}{2}}}{d})^2(1-\hat{p})\hat{p} \Leftrightarrow n_0 = (\frac{1{,}96}{0{,}05})^2 \cdot 0{,}25 \cong 384$
- Αφού $\frac{n_0}{N} = \frac{384}{750} = 0{,}512 > 0{,}05$ τότε
- $n = \frac{n_0}{1+\frac{n_0-1}{N}} \cong 255$ μέγεθος δείγματος για την παραπάνω διερεύνηση