Ανάλυση ποσοτικών δεδομένων (Ι)

Ε. Σαμόλη
Καθηγήτρια Επιδημιολογίας και Ιατρικής Στατιστικής
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
esamoli@med.uoa.gr

Βάνα Σύψα
Καθηγήτρια Επιδημιολογίας και Ιατρικής Στατιστικής
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
vsipsa@med.uoa.gr

Εκκρεμότητες από προηγούμενο μάθημα περιγραφικής στατιστικής

Περιγραφική Στατιστική Δεδομένων

Κατά προσέγγιση κανονικές κατανομές

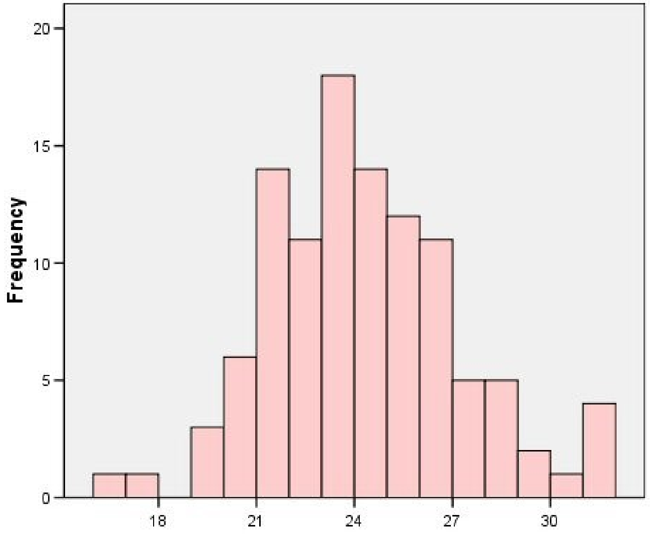
- Μέση Τιμή
- Σταθερή απόκλιση

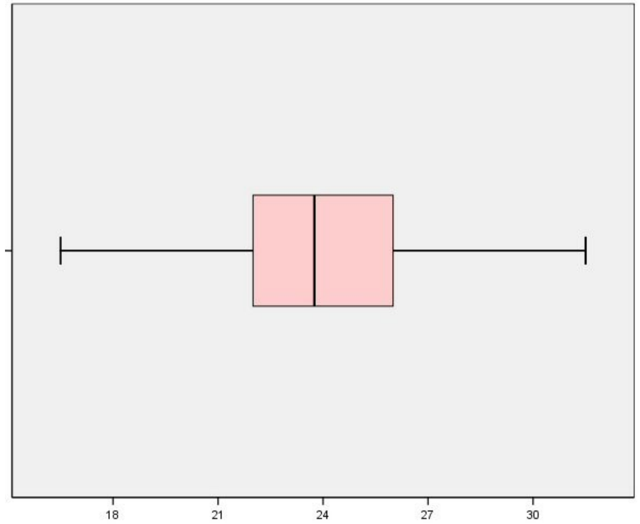
Ασύμμετρες κατανομές

- Διάμεσος
- Τεταρτημόρια (25ο, 75ο), Εκατοστημόρια (10ο,90ο)
- Ενδοτεταρτομοριακό εύρος (Interquartile range-IQR)

Διερεύνηση κανονικότητας μίας ποσοτικής μεταβλητής

- Ιστόγραμμα - Θηκόγραμμα





- Περιγραφικά δεδομένα

  → π.χ. μέση τιμή σε σύγκριση με τη διάμεση

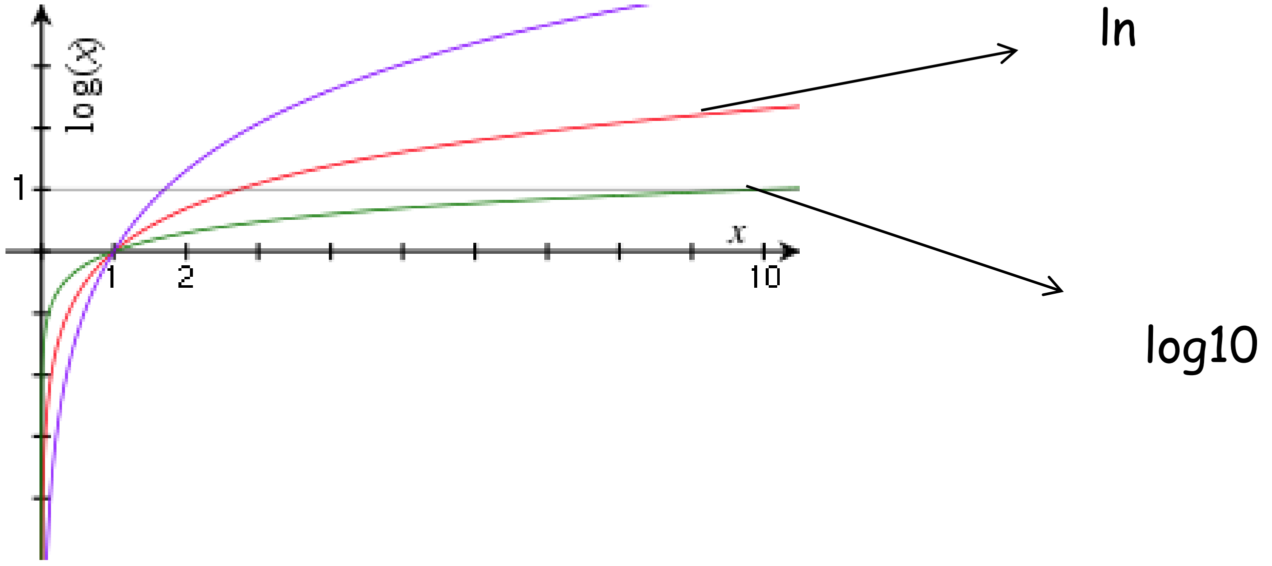
Τι κάνουμε αν τα δεδομένα δεν κατανέμονται κανονικά;

Μετασχηματισμός δεδομένων

- Όταν μία μεταβλητή δεν ακολουθεί την κανονική κατανομή → μετά από μετασχηματισμό μπορεί να προσεγγίζει την κανονική κατανομή

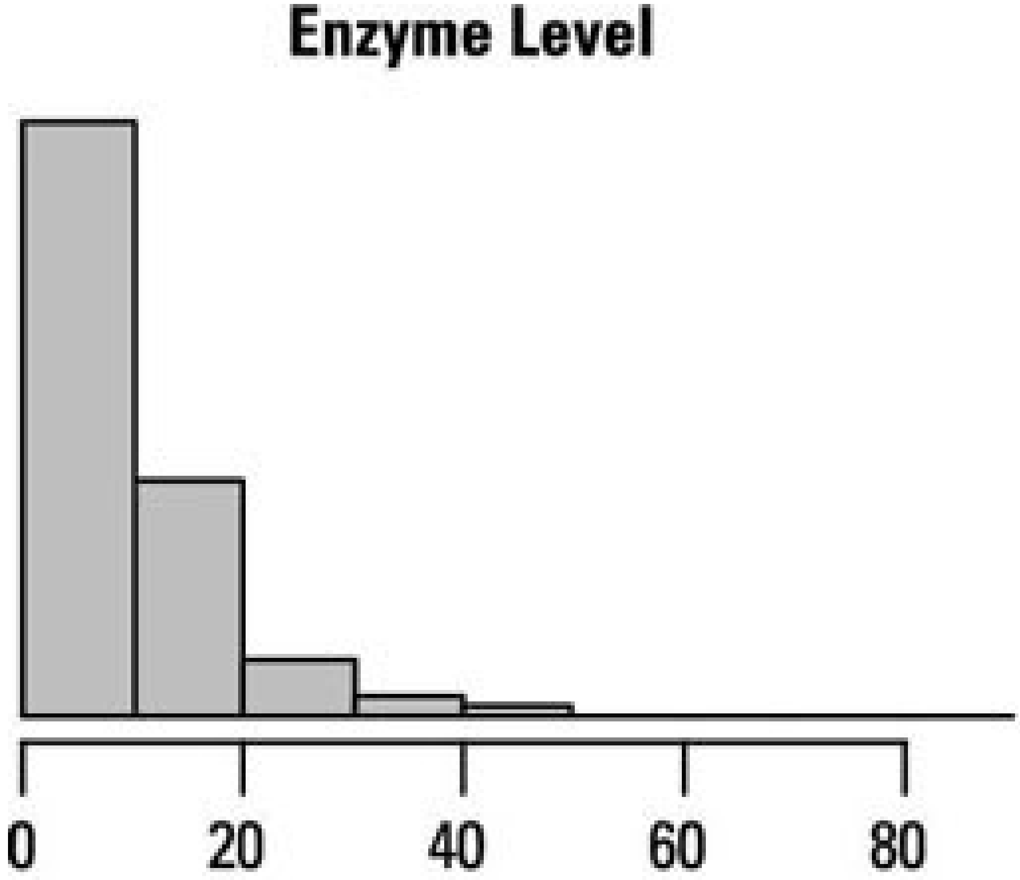
- Μετασχηματισμός
  - Π.χ. λογαριθμικός μετασχηματισμό z=log(y) για θετικά λοξές κατανομές (ln, $\log_{10}$)

Λογαριθμική συνάρτηση για x>0

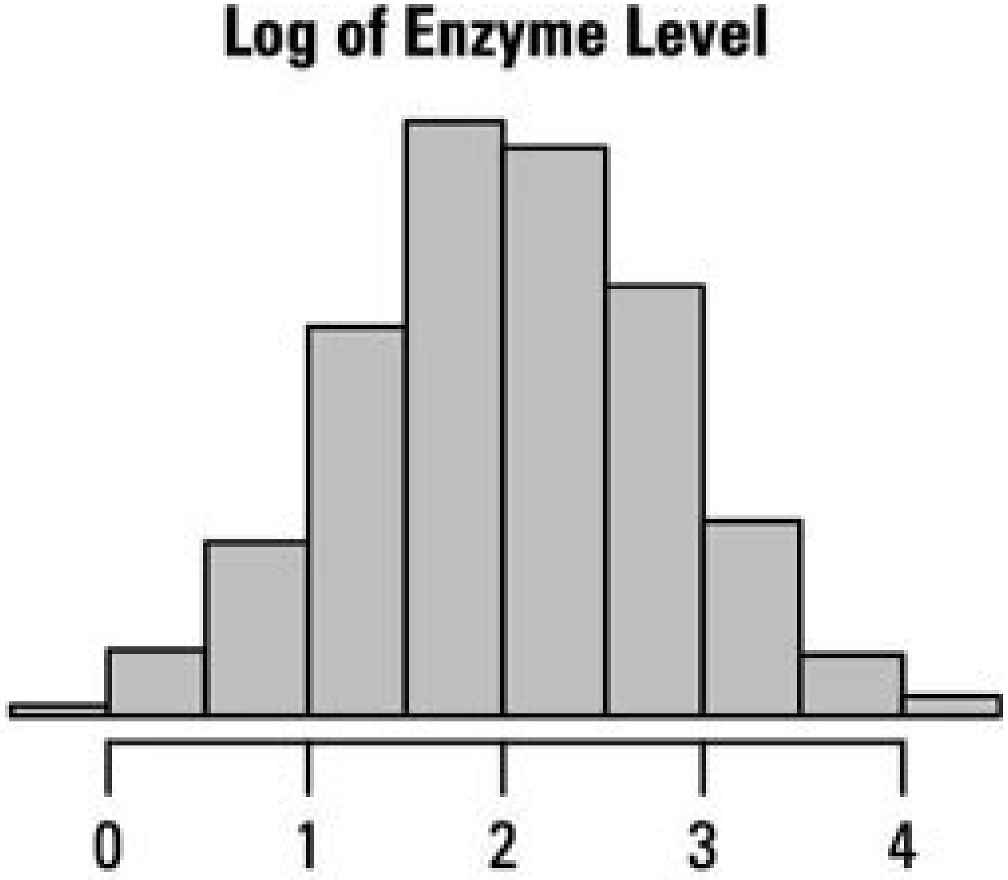


Μεταβλητή	x_1	x_2	x_3	x_4	x_5
Αρχικές τιμές	10	100	1.000	10.000	100.000
log10	1	2	3	4	5

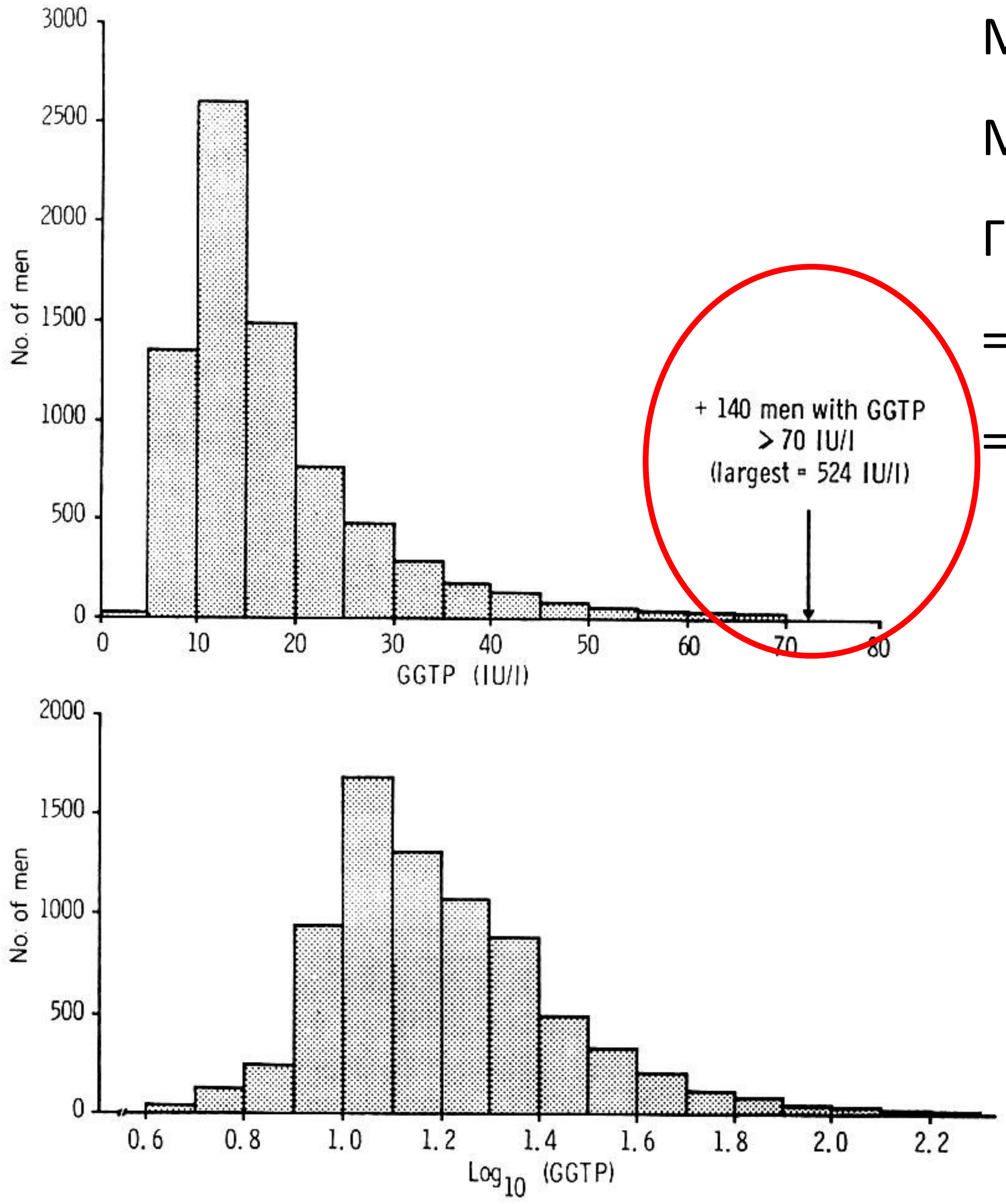
Παράδειγμα



a



b



Μέτρηση GGTP σε 7.613 άντρες

Μέση τιμή 19,2 IU/lt

Γεωμετρικός μέσος

=αντιλογάριθμος (mean(log))

= 15.6 IU/lt

Πηγή: Shaper et al. 1983

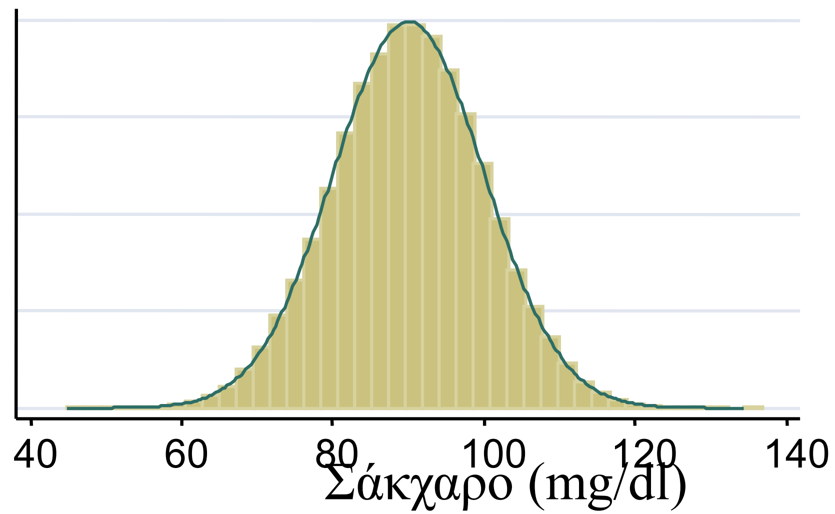
Προσδιορισμός φυσιολογικών τιμών ενός ποσοτικού μεγέθους

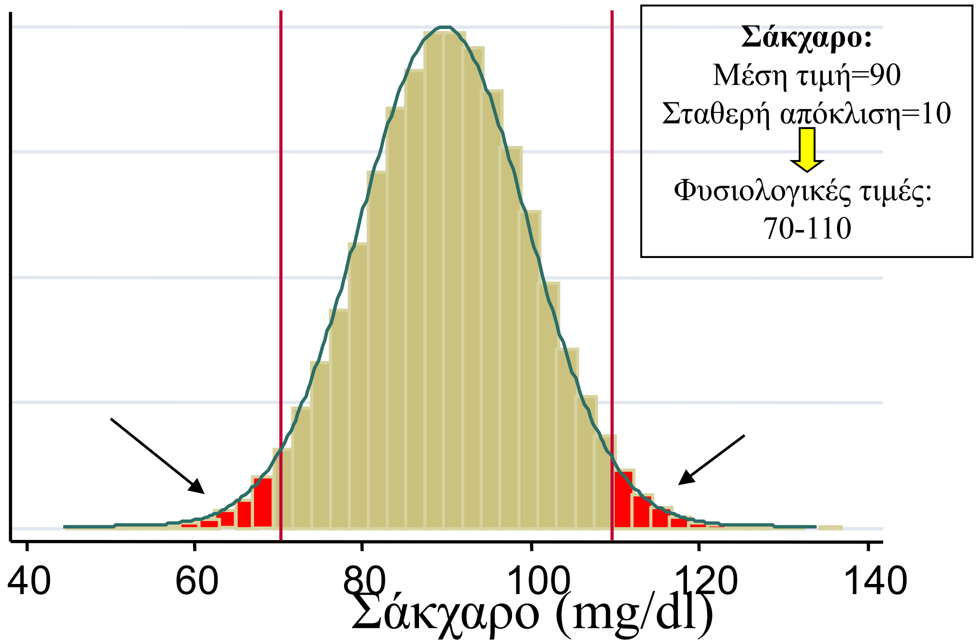
Φυσιολογικές τιμές

- Διάκριση παθολογικού – φυσιολογικού

Π.χ. Χοληστερόλη, δείκτης μάζας σώματος

Αρ. ατόμου	Επίπεδα σακχάρου
1	95
2	65
3	120
4	110
…	…
…	…
1000	85





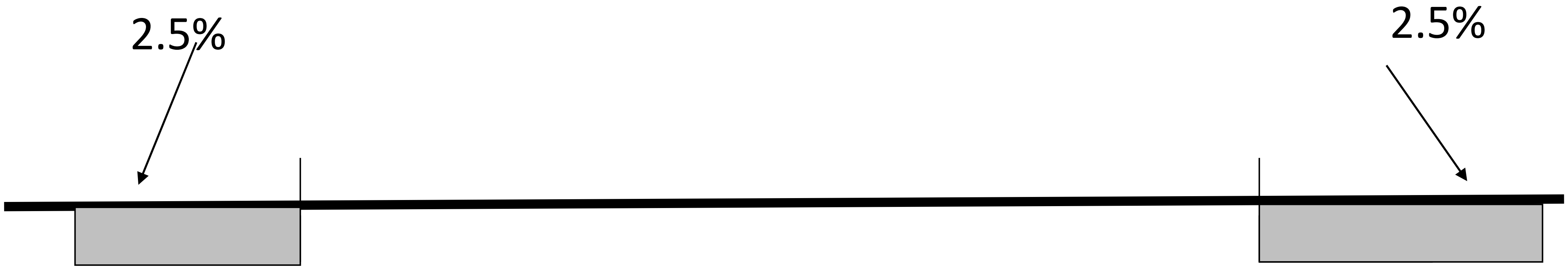
Όρια φυσιολογικών τιμών: Μέση τιμή ± 2 σταθερές αποκλίσεις → 95% των ατόμων

$$\overline{x} \pm 2 * SD$$

Για Ασύμμετρες (μη κανονικές) κατανομές

Φυσιολογικές τιμές → Το 95% του πληθυσμού να ανήκει μέσα στο διάστημα

Εκατοστημόρια (2.5, 97.5)



Υπολογισμός φυσιολογικών τιμών με στατιστικά κριτήρια

Όχι προσδιοριστικό →

πιθανολογικό χαρακτήρα

Οποιαδήποτε τιμή μπορεί να ανήκει σε υγιές άτομο, αλλά η **πιθανότητα** είναι **μικρότερη**, όταν η **απόσταση** από μέση τιμή είναι **μεγαλύτερη.**

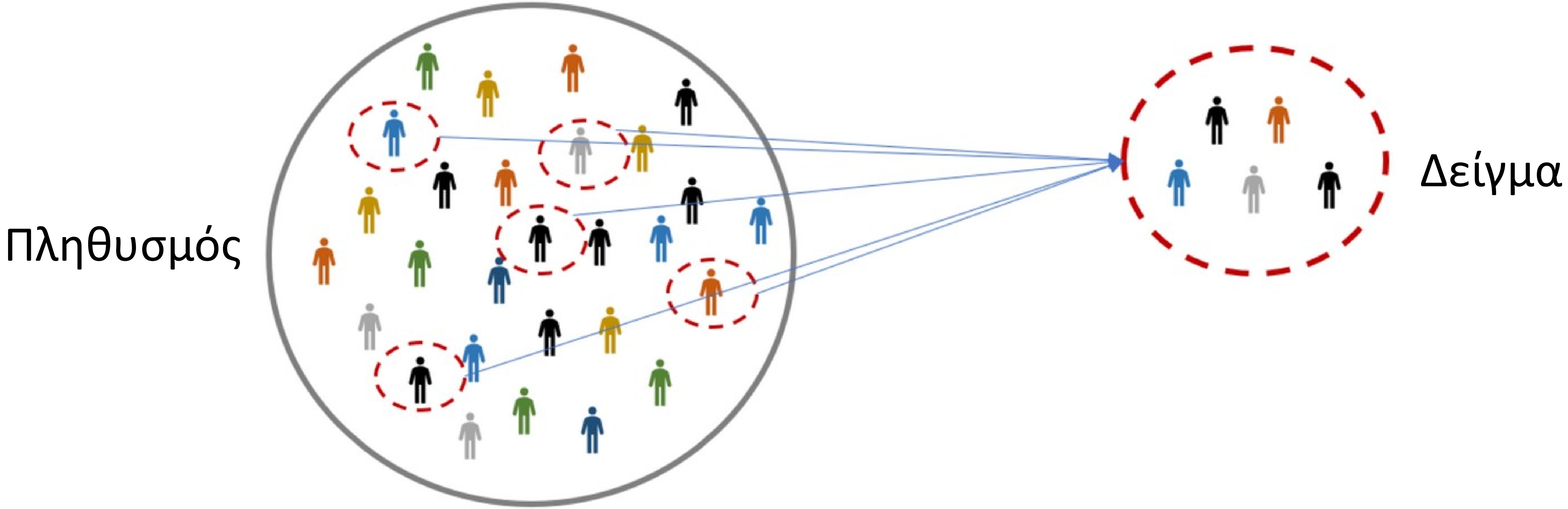
Ανάλυση ποσοτικών δεδομένων

Πιθανά ερωτήματα προς διερεύνηση

1. Ποιο είναι το ύψος των αγοριών ηλικίας 6 ετών στην Ελλάδα;
2. Διαφέρει το ύψος των αγοριών σήμερα σε σχέση με το αντίστοιχο ύψος το 1981;
3. Διαφέρει το ύψος των αγοριών στην Ελλάδα σε σχέση με το ύψος των κοριτσιών;
4. Μεταβάλλεται σημαντικά το ύψος των αγοριών από την ηλικία των 15 στην ηλικία των 17 ετών;

Πληθυσμός και δείγμα

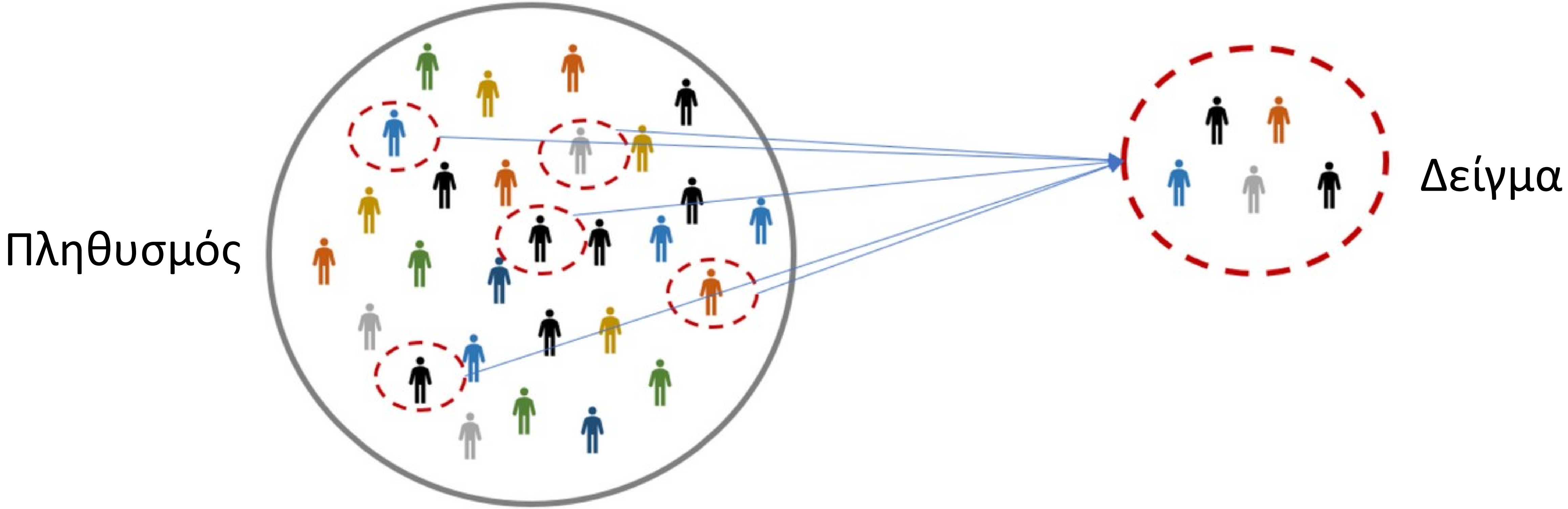
- Είναι δύσκολο να εξετάσουμε όλο τον πληθυσμό για να απαντήσουμε σε αυτά τα ερωτήματα
- Στην πράξη, επιλέγουμε ένα δείγμα από τον πληθυσμό και διερευνούμε τα ερωτήματα με τα δεδομένα από το δείγμα



Συμβολισμός στον πληθυσμό και το δείγμα

- **Πληθυσμός**: ελληνικοί χαρακτήρες
  - μέση τιμή → μ
  - σταθερή απόκλιση → σ

- **Δείγμα**: λατινικοί χαρακτήρες
  - μέση τιμή → $\bar{X}$
  - σταθερή απόκλιση → S



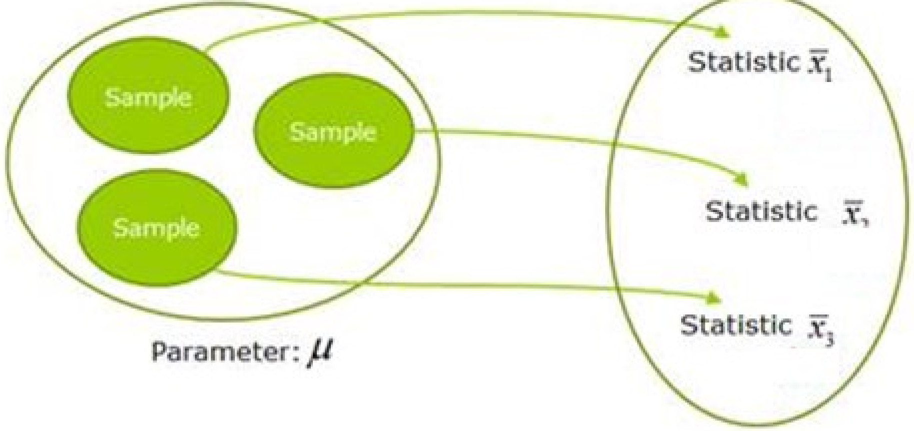
Πιθανά ερωτήματα προς διερεύνηση

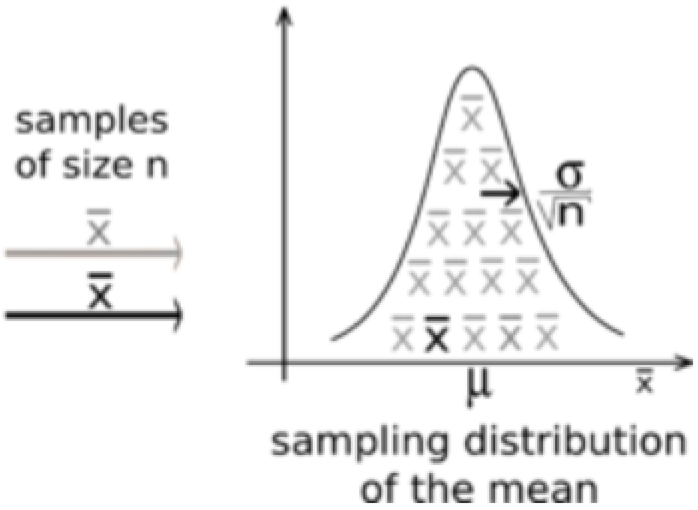
1. **Ποιο είναι το ύψος των αγοριών ηλικίας 6 ετών στην Ελλάδα;**
2. Διαφέρει το ύψος των αγοριών σήμερα σε σχέση με το αντίστοιχο ύψος το 1981;
3. Διαφέρει το ύψος των αγοριών στην Ελλάδα σε σχέση με το ύψος των κοριτσιών;
4. Μεταβάλλεται σημαντικά το ύψος των αγοριών από την ηλικία των 15 στην ηλικία των 17 ετών;

Πληθυσμός και δείγμα

- Επιλέγουμε ένα δείγμα και μετράμε το ανάστημα των παιδιών
- Είναι απίθανο η μέση τιμή που προκύπτει από το δείγμα που έχουμε επιλέξει να είναι ίση με τη μέση τιμή του πληθυσμού
- Π.χ. η επιλογή ενός άλλου δείγματος θα οδηγούσε σε άλλη μέση τιμή – ας υποθέσουμε ότι παίρνουμε πολλά δείγματα

Δειγματικοί μέσοι





[προκύπτει από το κεντρικό οριακό θεώρημα το οποίο δεν θα συζητήσουμε περαιτέρω]

- **Άρα, κατά μέσο όρο η μέση τιμή των δειγματικών μέσων $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ … θα είναι ίση με τη μέση τιμή μ στον πληθυσμού (αμερόληπτη εκτιμήτρια)**

Παράδειγμα

- Έστω οι τιμές ενός ποσοτικού μεγέθους σε ένα πληθυσμό με 5 άτομα (N=5) και μέση τιμή στον πληθυσμό: 116.8
- Παίρνουμε δείγματα των 3 ατόμων από τον πληθυσμό και υπολογίζουμε τους δειγματικούς μέσους

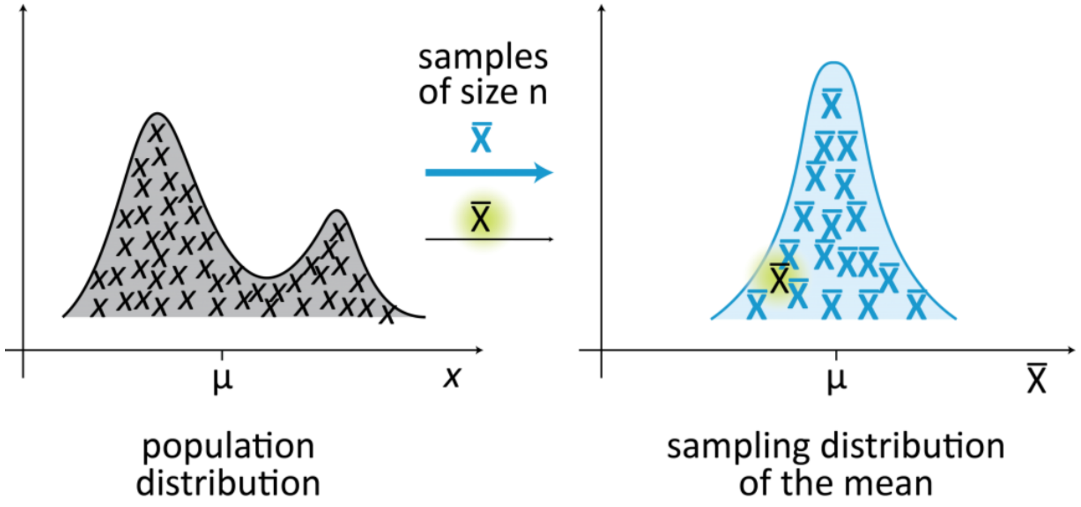
	Πληθυσμός
	100
	110
	118
	124
	132
Μέση τιμή	**116.8**

Δείγμα 1	**Δείγμα 2**	**Δείγμα 3**	**Δείγμα 4**	**Δείγμα 5**	**Δείγμα 6**	**Δείγμα 7**	**Δείγμα 8**	**Δείγμα 9**	**Δείγμα 10**
100	100	100	100	100	100				
110	110	110				110	100	110	
118			118	118		118	118		118
	124		124		124	124		124	124
		132		132	132		132	132	132
109.3	111.3	114.0	114.0	116.7	118.7	117.3	116.7	122.0	124.7

Μέση τιμή των μέσων από τα 10 δείγματα		**116.5**

Δειγματική κατανομή της μέσης τιμής

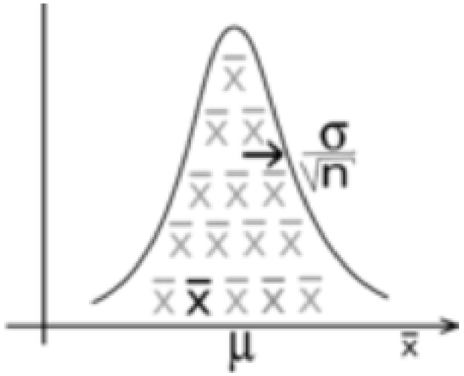
- Έστω ένα χαρακτηριστικό με **μέση τιμή μ** και **σταθερή απόκλιση σ**
- Αν έπαιρνα πολλαπλά δείγματα μεγέθους n από τον πληθυσμού και υπολόγιζα τη δειγματική μέση τιμή του χαρακτηριστικού που με ενδιαφέρει σε κάθε ένα από αυτά, θα μπορούσα να παραστήσω γραφικά την κατανομή τους



- Από τη θεωρία, σε μεγάλα δείγματα, η δειγματική κατανομή των μέσων τιμών ακολουθεί την κανονική κατανομή με τα εξής χαρακτηριστικά:
    - Η μέση τιμή της είναι ίση με τη μέση τιμή μ στον πληθυσμό
    - Η σταθερή απόκλιση των δειγματικών μέσων είναι ίση με $\frac{\sigma}{\sqrt{n}}$ → **Τυπικό σφάλμα της μέσης τιμής (Standard Error ή SE)**

Τι εκφράζει το τυπικό σφάλμα της μέσης τιμής (Standard error, SE):

Είναι μέτρο της **απόστασης** της **μέσης τιμής του δείγματος** από την **πραγματική μέση τιμή** στον πληθυσμό

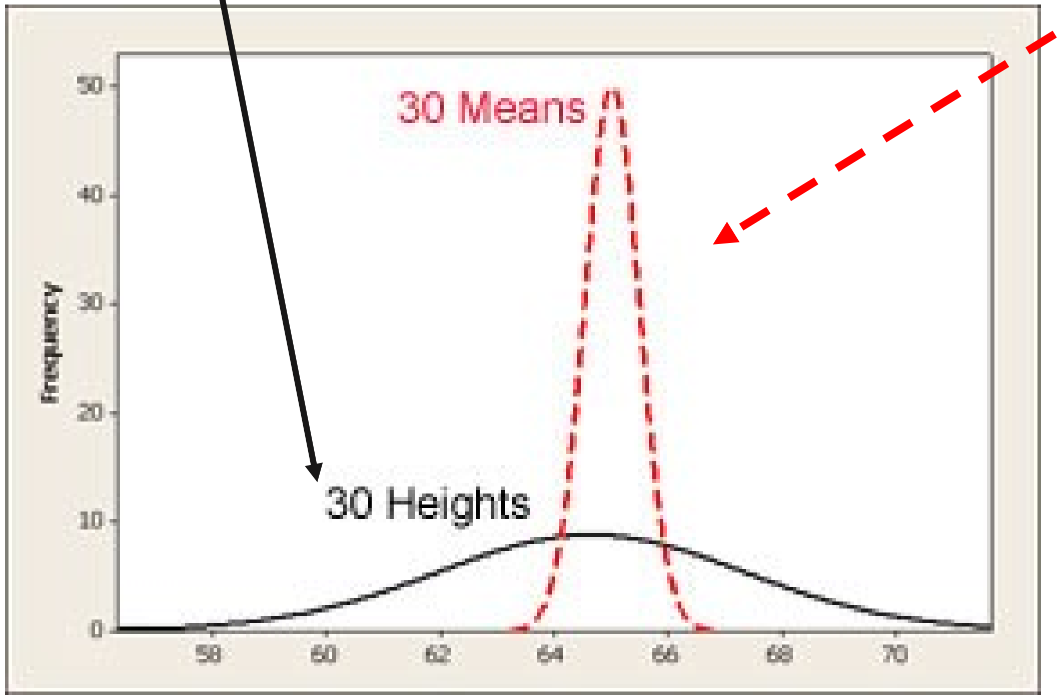


$$SE = \frac{\sigma}{\sqrt{n}}$$

- Το μέγεθος του εξαρτάται από:
  1) Την ετερογένεια που έχει το χαρακτηριστικό που μελετάμε στον πληθυσμό
  2) Το μέγεθος του δείγματος
- Όσο πιο μεγάλο το μέγεθος του δείγματος n, τόσο πιο μικρό το τυπικό σφάλμα
- Επειδή συνήθως δεν ξέρουμε το σ, το SE υπολογίζεται ως $\boxed{SE = \frac{S}{\sqrt{n}}}$

Προσοχή! Το Standard Error (SE) διαφέρει από τη Standard Deviation (SD)

- Π.χ. Αν πάρω ένα δείγμα 30 ατόμων και μετρήσω το ύψος, η σταθερή απόκλιση **εκφράζει την ετερογένεια στα άτομα**



- Αν πάρω 30 δείγματα ατόμων και υπολογίσω τους δειγματικούς μέσους, το τυπικό σφάλμα εκφράζει με πόση ακρίβεια εκτιμάται η μέση τιμή στον πληθυσμό από τους δειγματικούς μέσους
  - Στην πράξη το standard error είναι η σταθερή απόκλιση των δειγματικών μέσων (**εκφράζει την ετερογένεια στους δειγματικούς μέσους**)

SD	SE
$s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}}$	$SE = \dfrac{S}{\sqrt{n}}$
Εκφράζει τη διασπορά των παρατηρήσεων	Εκφράζει την απόσταση του δειγματικού μέσου από τον πραγματικό μέσο
Δεν επηρεάζεται ιδιαίτερα από το μέγεθος του δείγματος	Επηρεάζεται από το μέγεθος του δείγματος n

Ερώτημα: Πώς μπορούμε να εκτιμήσουμε τη μέση τιμή ενός μεγέθους στον πληθυσμό από ένα δείγμα;

- Αναφέραμε ότι αν πάρουμε πολλαπλά δείγματα και υπολογίσουμε τη μέση τιμή σε κάθε ένα από αυτά, ο μέσος όρος των δειγματικών μέσων τιμών προσεγγίζει τη μέση τιμή του χαρακτηριστικού στον πληθυσμό

- **ΌΜΩΣ, στην πράξη δεν μπορούμε να μετρήσουμε όλο τον πληθυσμό ούτε να μετρήσουμε πολλαπλά δείγματα του πληθυσμού**

→ πρέπει να βγάλουμε συμπέρασμα από το ένα δείγμα που μελετάμε και την εκτίμηση της μέσης τιμής $\boldsymbol{\bar{X}}$ και της σταθερής απόκλισης *S* από το δείγμα αυτό

- Εκτιμούμε το δειγματικό μέσο $\boldsymbol{\bar{X}}$ & συνοδεύουμε την εκτίμηση αυτή με ένα διάστημα εντός του οποίου είμαστε σχεδόν σίγουροι ότι βρίσκεται η πραγματική μέση τιμή **μ**
    - Σχεδόν σίγουροι: 95% ή 99% σίγουροι → **95% ή 99% διάστημα εμπιστοσύνης**

Διάστημα εμπιστοσύνης μέσης τιμής

- Πώς υπολογίζεται :

Η εκτίμησή μας για τη μέση τιμή $\pm$ περιθώριο σφάλματος

$\bar{X}$

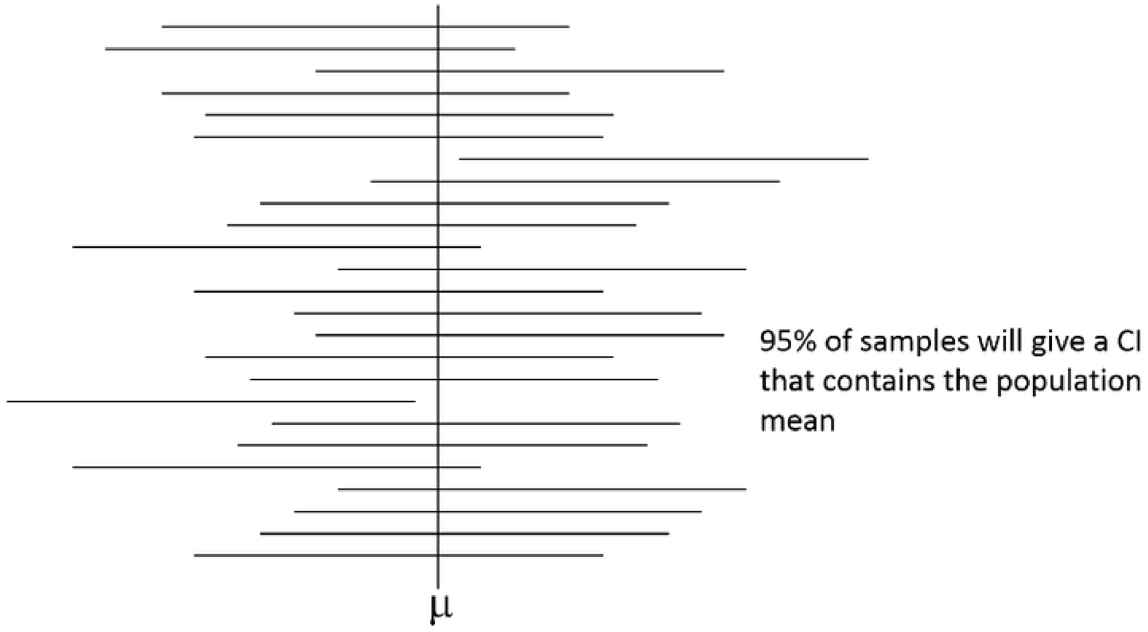
Καθορίζεται από το πόσο σίγουροι θέλουμε να είμαστε (95%, 99% κλπ), το μέγεθος του δείγματος και την ετερογένεια του χαρακτηριστικού στον πληθυσμό

	Για μεγάλο δείγμα (κανονική κατανομή)	Για μικρό δείγμα (S όχι αξιόπιστη εκτίμηση της σ) (κατανομή t)
95% διάστημα εμπιστοσύνης	$\bar{X} \pm 1.96 \cdot SE$	$\bar{X} \pm t_{n-1,0.05} \cdot SE$
99% διάστημα εμπιστοσύνης	$\bar{X} \pm 2.58 \cdot SE$	$\bar{X} \pm t_{n-1,0.01} \cdot SE$

t: θα το δούμε σε επόμενη διαφάνεια

Τι εκφράζει το 95% διάστημα εμπιστοσύνης;

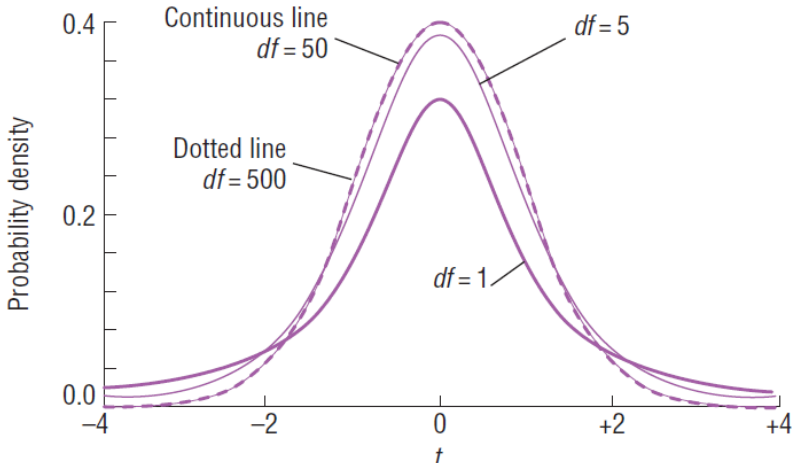
- Αν παίρναμε 100 διαφορετικά δείγματα και υπολογίζαμε το 95% διάστημα εμπιστοσύνης σε κάθε ένα, τότε τα 95 από αυτά θα περιλάμβαναν τον πληθυσμιακό μέσο



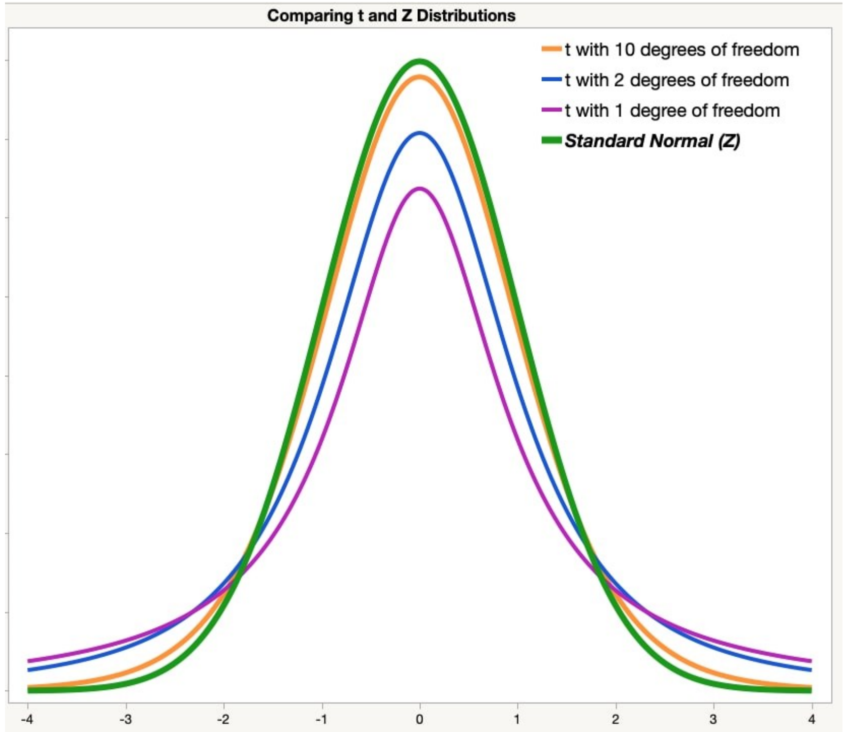
- Συνήθως αναφέρουμε ότι υπάρχει 95% πιθανότητα η πραγματική μέση τιμή να είναι εντός αυτού του διαστήματος (όχι εντελώς ακριβές αλλά θα το χρησιμοποιούμε)

Λίγα λόγια για την κατανομή t

- Συμμετρική κωδωνοειδής κατανομή – έχει πιο «παχιές» άκρες από την κανονική κατανομή
- Το σχήμα της εξαρτάται από τους βαθμούς ελευθερίας (degrees of freedom) → df = n-1



Για μεγάλο μέγεθος δείγματος η κατανομή t προσεγγίζει την κανονική



Παράδειγμα εκτίμησης μέσης τιμής και 95% διαστήματος εμπιστοσύνης

- Θέλουμε να εκτιμήσουμε το ανάστημα των αγοριών ηλικίας 6 ετών στην Ελλάδα

- **Πληθυσμός**: όλα τα αγόρια ηλικίας 6 ετών στην Ελλάδα
  → δεν μπορούμε να κάνουμε μέτρηση σε όλα!

- **Δείγμα**: 730 αγόρια ηλικίας 6 ετών (Smpokos et al, 2019)
  - Μέση τιμή (SD): 119.6 (6.1) cm

Υπολογισμός τυπικού σφάλματος (SE) και 95% διαστήματος εμπιστοσύνης (95% ΔΕ)

- $\boxed{SE = \frac{S}{\sqrt{n}} = \frac{6.1}{\sqrt{730}} = 0.226}$

- 95% ΔΕ (μεγάλο δείγμα): $\bar{X} \pm 1.96 \cdot SE =$
  $= 119.6 \pm 1.96 * 0.226$ → (119.2, 120.0) cm

Ερμηνεία του αποτελέσματος

- Ποια/ποιες από τις παρακάτω προτάσεις είναι λάθος;

1. Αν παίρναμε 100 διαφορετικά δείγματα και υπολογίζαμε το 95% διάστημα εμπιστοσύνης σε κάθε ένα, τότε τα 95 από αυτά θα περιλάμβαναν τον πληθυσμιακό μέσο

   Σωστό

2. Είμαι 95% σίγουρος ότι η πραγματική μέση τιμή του ύψους των αγοριών ηλικίας 6 ετών βρίσκεται εντός του διαστήματος (119.2, 120.0)

   Όχι εντελώς ακριβής αλλά θα τη χρησιμοποιούμε

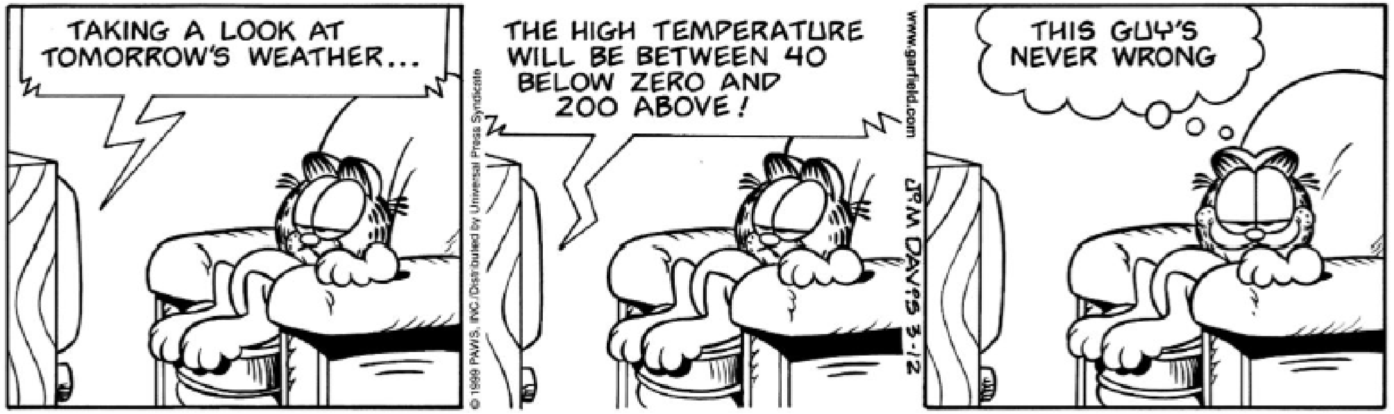
3. Το 95% των αγοριών ηλικίας 6 ετών έχουν ύψος μεταξύ 119.2 και 120.0 cm

   Λάθος

Μας ικανοποιεί το αποτέλεσμα;

- Από μία μελέτη σε 730 παιδιά, συμπεράναμε ότι το μέσο ύψος των εξάχρονων αγοριών στην Ελλάδα βρίσκεται με 95% πιθανότητα μεταξύ 119.2 και 120.0 cm

→ αρκετά «στενό» το εύρος



Άλλα ερωτήματα

- Αν το δείγμα που είχαμε μελετήσει ήταν μικρότερο, πώς θα ήταν το διάστημα εμπιστοσύνης; Πιο ευρύ ή πιο στενό;
- Αν θέλαμε το 99% ΔΕ, πώς θα ήταν; Πιο ευρύ ή πιο στενό;
- Αν είχαμε μελετήσει όλο τον πληθυσμό, πώς φαντάζεστε το διάστημα εμπιστοσύνης;

Παράδειγμα με μικρό μέγεθος δείγματος

- Μέση τιμή (SD): 119.6 (6.1) cm αλλά **μέγεθος δείγματος n=20 άτομα**

- 95% ΔΕ: $\bar{X} \pm t_{n-1,0.05} \cdot SE$

- $t_{20-1,0.05}=2.093$

- $SE = \dfrac{S}{\sqrt{n}} = \dfrac{6.1}{\sqrt{20}} = 1.36$

- 95% ΔΕ: $119.6 \pm 2.09 \cdot 1.36$ → 116.7, 122.5

Άρα **95% ΔΕ → (116.7, 122.5)**

99% Διάστημα εμπιστοσύνης;

t-distribution.

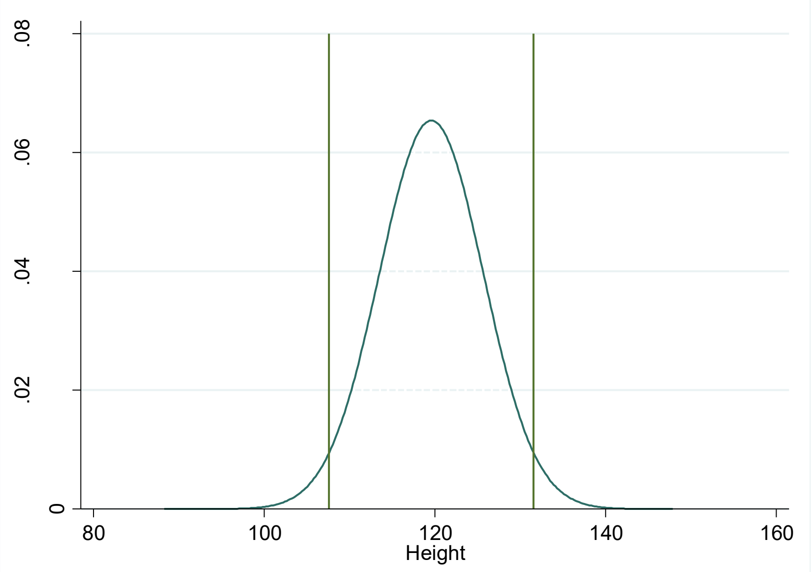
	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	6.314	12.706	63.656	636.58
2	2.920	4.303	9.925	31.600
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850

95% Διάστημα εμπιστοσύνης και φυσιολογικές τιμές

Συχνά γίνεται σύγχυση μεταξύ των φυσιολογικών τιμών και του διαστήματος εμπιστοσύνης της μέσης τιμής

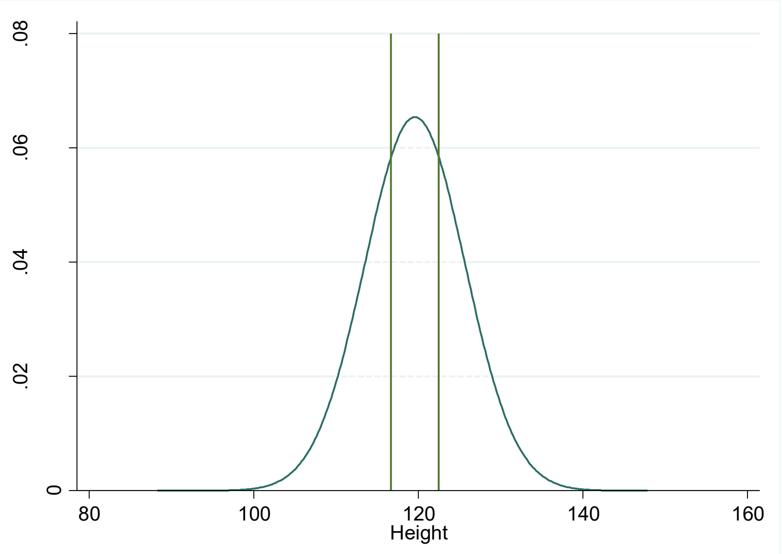
$\bar{X} \pm 1.96*SD$

Εκφράζει το εύρος που βρίσκεται το 95% των παρατηρήσεων



$\bar{X} \pm 1.96*SE$

Εκφράζει το εύρος που βρίσκεται το 95% των δειγματικών μέσων



Παράδειγμα: Ανάστημα εφήβων ηλικίας 17 ετών στην Ελλάδα

J PREV MED HYG 2018; 59: E36-E47

ORIGINAL ARTICLE

Current data in Greek children indicate decreasing trends of obesity in the transition from childhood to adolescence; results from the National Action for Children's Health (EYZHN) program

K.D. TAMBALIS[1], D.B. PANAGIOTAKOS[1], G. PSARRA[1], L.S. SIDOSSIS[1 2]
[1] Department of Nutrition and Dietetics, Harokopio University, Athens, Greece; [2] Department of Kinesiology and Health, Rutgers University, New Brunswick, USA

Παράδειγμα: Ανάστημα εφήβων ηλικίας 17 ετών στην Ελλάδα

	Boys		Girls	
Age †	N	Height (cm)	N	Height (cm)
17	2358	177.0 (7.1)*	2228	164.7 (6.3)

Με βάση τα παραπάνω δεδομένα:

1. Σε ποιο εύρος κυμαίνεται το ύψος του 95% των αγοριών; Σε ποιο εύρος κυμαίνεται το ύψος του 95% των κοριτσιών;
2. Ποιο είναι το 95% διάστημα εμπιστοσύνης της μέσης τιμής ύψους σε αγόρια και κορίτσια;

	95% ΔΕ της μέσης τιμή	Εύρος εντός του οποίου κυμαίνεται το 95% του ύψους των παιδιών
Αγόρια		
Κορίτσια		

Παράδειγμα: Ανάστημα εφήβων ηλικίας 17 ετών στην Ελλάδα

	Boys		Girls	
Age †	N	Height (cm)	N	Height (cm)
17	2358	177.0 (7.1)*	2228	164.7 (6.3)

Με βάση τα παραπάνω δεδομένα:

1. Σε ποιο εύρος κυμαίνεται το ύψος του 95% των αγοριών; Σε ποιο εύρος κυμαίνεται το ύψος του 95% των κοριτσιών;
2. Ποιο είναι το 95% διάστημα εμπιστοσύνης της μέσης τιμής ύψους σε αγόρια και κορίτσια;

	95% ΔΕ της μέσης τιμή	Εύρος εντός του οποίου κυμαίνεται το 95% του ύψους των παιδιών
Αγόρια	(176.7, 177.3)	(163.1, 190.9)
Κορίτσια	(164.4, 165.0)	(152.4, 177.0)

Πιθανά ερωτήματα προς διερεύνηση

1. Ποιο είναι το ύψος των αγοριών ηλικίας 6 ετών στην Ελλάδα;
2. **Διαφέρει το ύψος των αγοριών σήμερα σε σχέση με το αντίστοιχο ύψος το 1981;**
3. **Διαφέρει το ύψος των αγοριών στην Ελλάδα σε σχέση με το ύψος των κοριτσιών;**
4. **Μεταβάλλεται σημαντικά το ύψος των αγοριών από την ηλικία των 15 στην ηλικία των 17 ετών;**

→ Έλεγχος υποθέσεων

Έλεγχος υποθέσεων

Έλεγχος υποθέσεων

- Και σε αυτή την περίπτωση, καλούμαστε να δώσουμε απάντηση σε ένα ερώτημα που αφορά τον πληθυσμό χρησιμοποιώντας τα δεδομένα από ένα δείγμα

- Όταν χρησιμοποιούμε περιγραφική στατιστική σε δεδομένα που έχουμε συλλέξει σε ένα δείγμα, μπορεί να δούμε π.χ. διαφορές σε ένα μέγεθος μεταξύ 2 συγκρινόμενων ομάδων
  - Οι διαφορές αυτές αντανακλούν **πραγματική διαφορά ή είναι αποτέλεσμα τύχης;**

Τα βήματα στον έλεγχο υποθέσεων

1. Διατυπώνουμε το ερώτημα και καθορίζουμε τη μηδενική και την εναλλακτική υπόθεση
   - Μηδενική υπόθεση H_0: δεν υπάρχει διαφορά
   - Εναλλακτική υπόθεση H_1: υπάρχει διαφορά
2. Συλλέγουμε δεδομένα από ένα δείγμα
3. Υπολογίζουμε την τιμή της <u>κατάλληλης</u> στατιστικής συνάρτησης με βάση τα δεδομένα από το δείγμα
4. Συγκρίνουμε την τιμή της στατιστικής συνάρτησης σε σχέση με τιμή από κατανομή που θα αναμέναμε αν ίσχυε η μηδενική υπόθεση (αν δηλαδή δεν υπάρχει διαφορά)
5. Επιλέγουμε το επίπεδο σημαντικότητας α, δηλαδή την πιθανότητα σφάλματος τύπου I που είμαστε διατεθειμένοι να ανεχτούμε → **Σφάλμα τύπου I:** να συμπεράνουμε λανθασμένα ότι υπάρχει διαφορά ενώ στην πραγματικότητα δεν υπάρχει (η πιθανότητα να απορρίψουμε τη μηδενική απόφαση λανθασμένα) → α=5%
6. Με βάση την τιμή της στατιστικής συνάρτησης και το επίπεδο σημαντικότητας, απορρίπτουμε ή δεν απορρίπτουμε τη μηδενική υπόθεση

Μηδενική και εναλλακτική υπόθεση

- Π.χ. θέλουμε να διερευνήσουμε αν υπάρχει διαφορά στα επίπεδα αρτηριακής πίεσης μεταξύ καπνιστών και μη καπνιστών

H_0: μέσα επίπεδα πίεσης σε καπνιστές = μέσα επίπεδα σε μη καπνιστές

H_1: μέσα επίπεδα πίεσης σε καπνιστές ≠ μέσα επίπεδα σε μη καπνιστές

- Π.χ. θέλουμε να διερευνήσουμε αν υπάρχει διαφορά στο ποσοστό ανταπόκρισης μεταξύ ασθενών που έλαβαν το φάρμακο A σε σχέση με αυτούς που έλαβαν το φάρμακο B

H_0: % ανταπόκρισης στο φάρμακο A = % ανταπόκρισης στο φάρμακο B

H_1: % ανταπόκρισης στο φάρμακο A ≠ % ανταπόκρισης στο φάρμακο B

Στην εναλλακτική υπόθεση σχεδόν ποτέ δεν προσδιορίζουμε κάποια φορά

- π.χ. μέσα επίπεδα πίεσης σε καπνιστές> από τα μέσα επίπεδα σε μη καπνιστές
- Ελέγχουμε το ≠ (two-sided test / αμφίπλευρος έλεγχος)

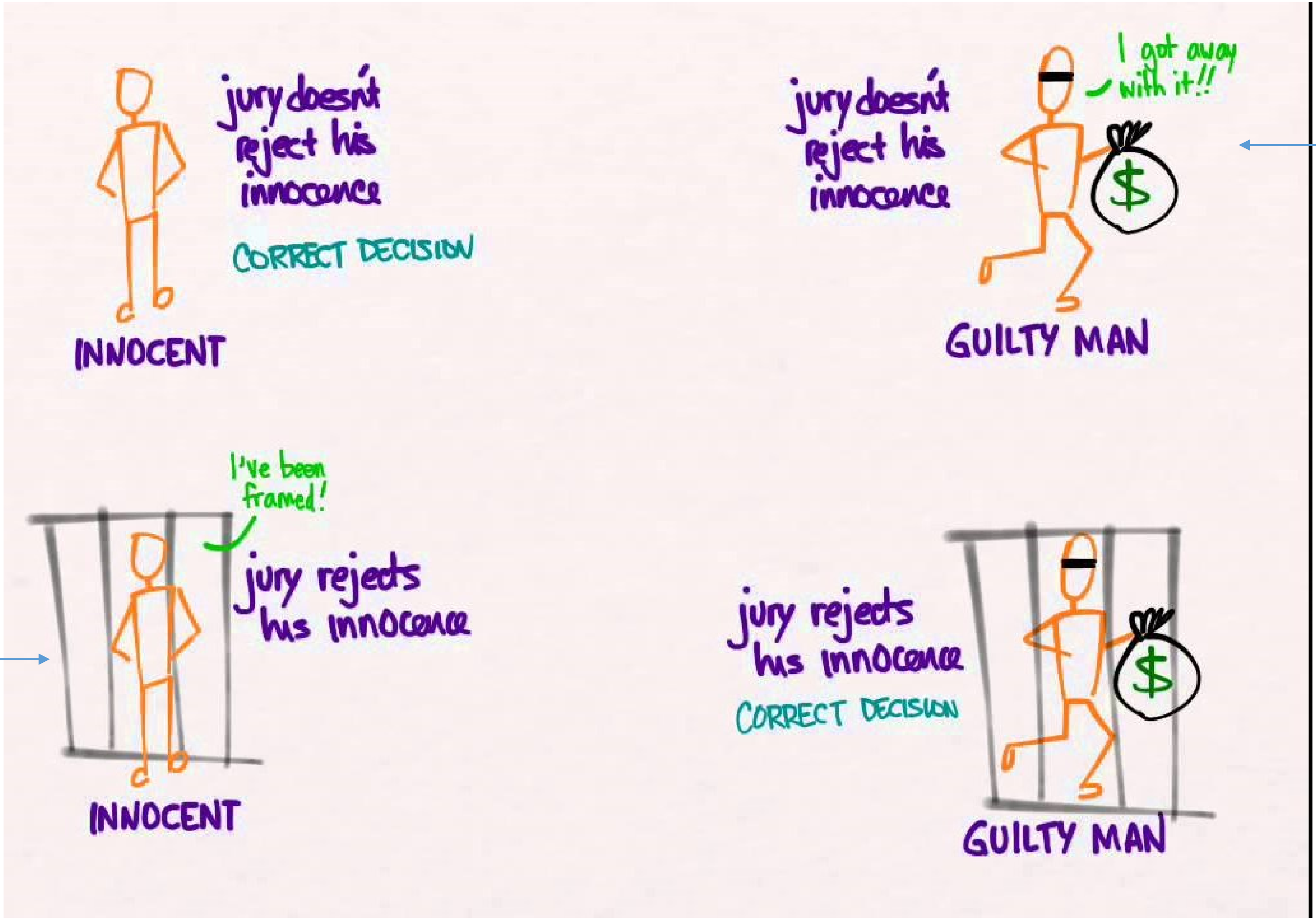
Η λογική του ελέγχου υποθέσεων

- Στη νομική επιστήμη, εφαρμόζεται το κριτήριο της αθωότητας για τον κατηγορούμενο έως αποδείξεως του εναντίου

  Μηδενική υπόθεση: ο κατηγορούμενος είναι αθώος

- Συλλέγονται δεδομένα προκειμένου να διερευνηθεί αν αυτά είναι συμβατά με αυτή την υπόθεση
  - Αν είναι συμβατά **δεν μπορώ να απορρίψω αυτή την υπόθεση** και αθωώνω τον κατηγορούμενο
  - Αν δεν είναι συμβατά, **απορρίπτω τη μηδενική υπόθεση** → ένοχος

- Σε αυτό τον έλεγχο μπορεί να υπεισέλθει κάποιο σφάλμα



Σφάλμα τύπου II

Σφάλμα τύπου I

Σφάλματα στον έλεγχο υποθέσεων

Αποτέλεσμα στο δείγμα της μελέτης	**Αλήθεια στον πληθυσμό**	
	Υπάρχει διαφορά	Δεν υπάρχει διαφορά
Υπάρχει διαφορά	**Σωστό (Ισχύς)**	**Σφάλμα τύπου I (a)**
Δεν υπάρχει διαφορά	**Σφάλμα τύπου II (β)**	**Σωστό**

Στατιστικά σφάλματα και ισχύς (& επιθυμητά επίπεδα)

α = σφάλμα τύπου Ι (επίπεδο σημαντικότητας)

= η πιθανότητα να ανιχνευθεί διαφορά όταν ΔΕΝ υπάρχει **ΣΥΝΗΘΩΣ 5%**

β = σφάλμα τύπου ΙΙ

= η πιθανότητα να ΜΗΝ ανιχνευθεί διαφορά όταν υπάρχει ΣΥΝΗΘΩΣ 10%-20%

1-β = ισχύς της μελέτης να ανιχνεύσει την αναμενόμενη διαφορά

= η πιθανότητα να ανιχνευθεί η αναμενόμενη διαφορά όταν υπάρχει (ΣΥΝΗΘΩΣ 80%-90%)

Στον έλεγχο υποθέσεων, ορίζουμε το επίπεδο σημαντικότητας στο 5%

Πιθανά ερωτήματα προς διερεύνηση

1. Ποιο είναι το ύψος των αγοριών ηλικίας 6 ετών στην Ελλάδα;
2. **Διαφέρει το ύψος των αγοριών σήμερα σε σχέση με το αντίστοιχο ύψος το 1981;**
3. Διαφέρει το ύψος των αγοριών στην Ελλάδα σε σχέση με το ύψος των κοριτσιών;
4. Μεταβάλλεται σημαντικά το ύψος των αγοριών από την ηλικία των 15 στην ηλικία των 17 ετών;

Παράδειγμα

- Εκτιμάται ότι το 1981, το μέσο ύψος των αγοριών ηλικίας 6 ετών ήταν 117.0 cm

- Το 2015, σε δείγμα 730 αγοριών ηλικίας 6 ετών εκτιμήθηκε για το ύψος:
    - Μέση τιμή (SD): 119.6 (6.1) cm (Smpokos et al, 2019)

- Παρατηρούμε διαφορά στις μέσες τιμές (117.0 vs. 119.6)

- Προέκυψε αυτή η διαφορά τυχαία ή μπορούμε να συμπεράνουμε ότι διαχρονικά το ύψος των αγοριών μεταβλήθηκε;

Έλεγχος της υπόθεσης αυτής

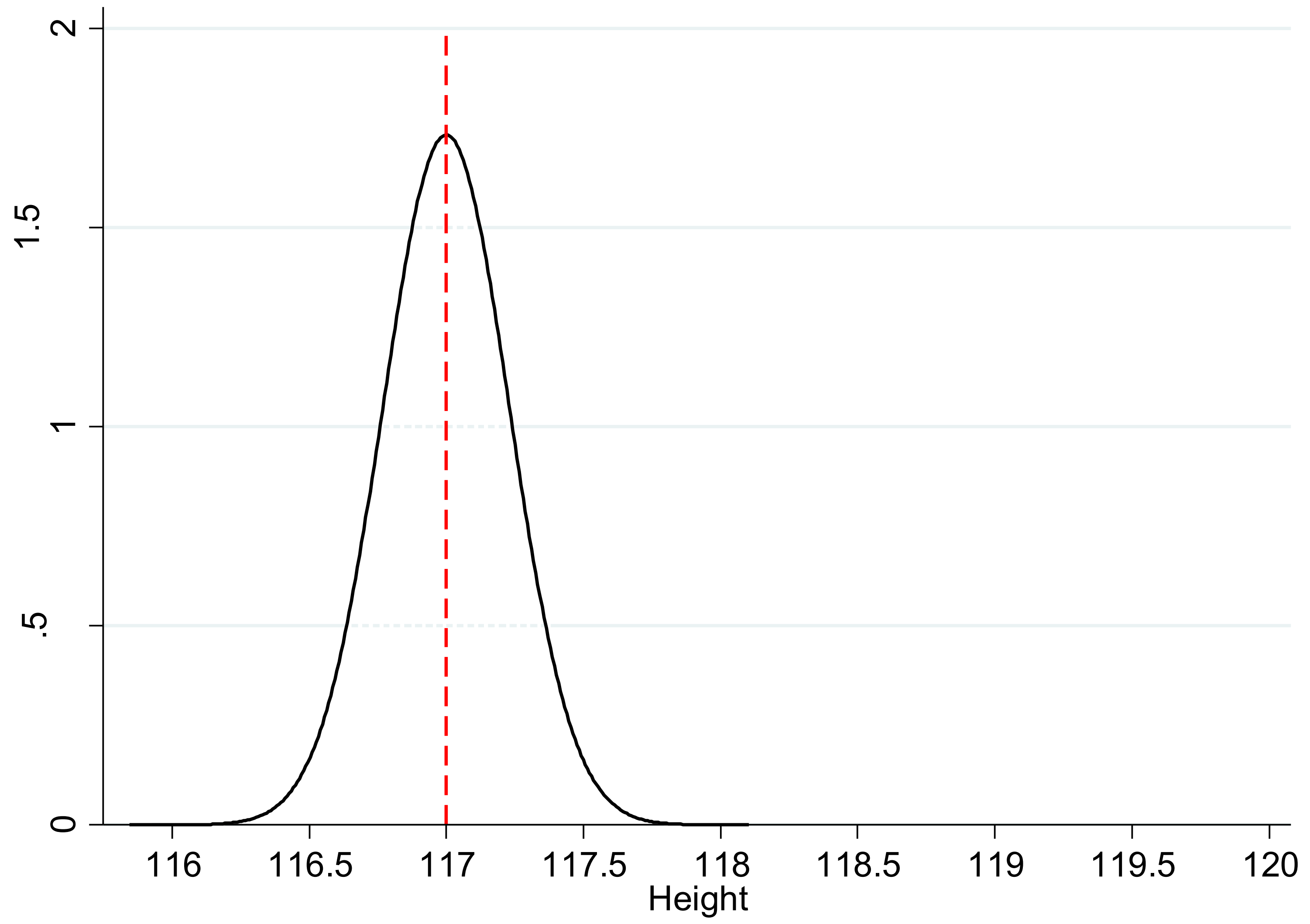
- Διαφέρει το μέσο ύψος των αγοριών ηλικίας 6 ετών το 2015 σε σχέση με την τιμή του 1981;

H0: μ=117.0

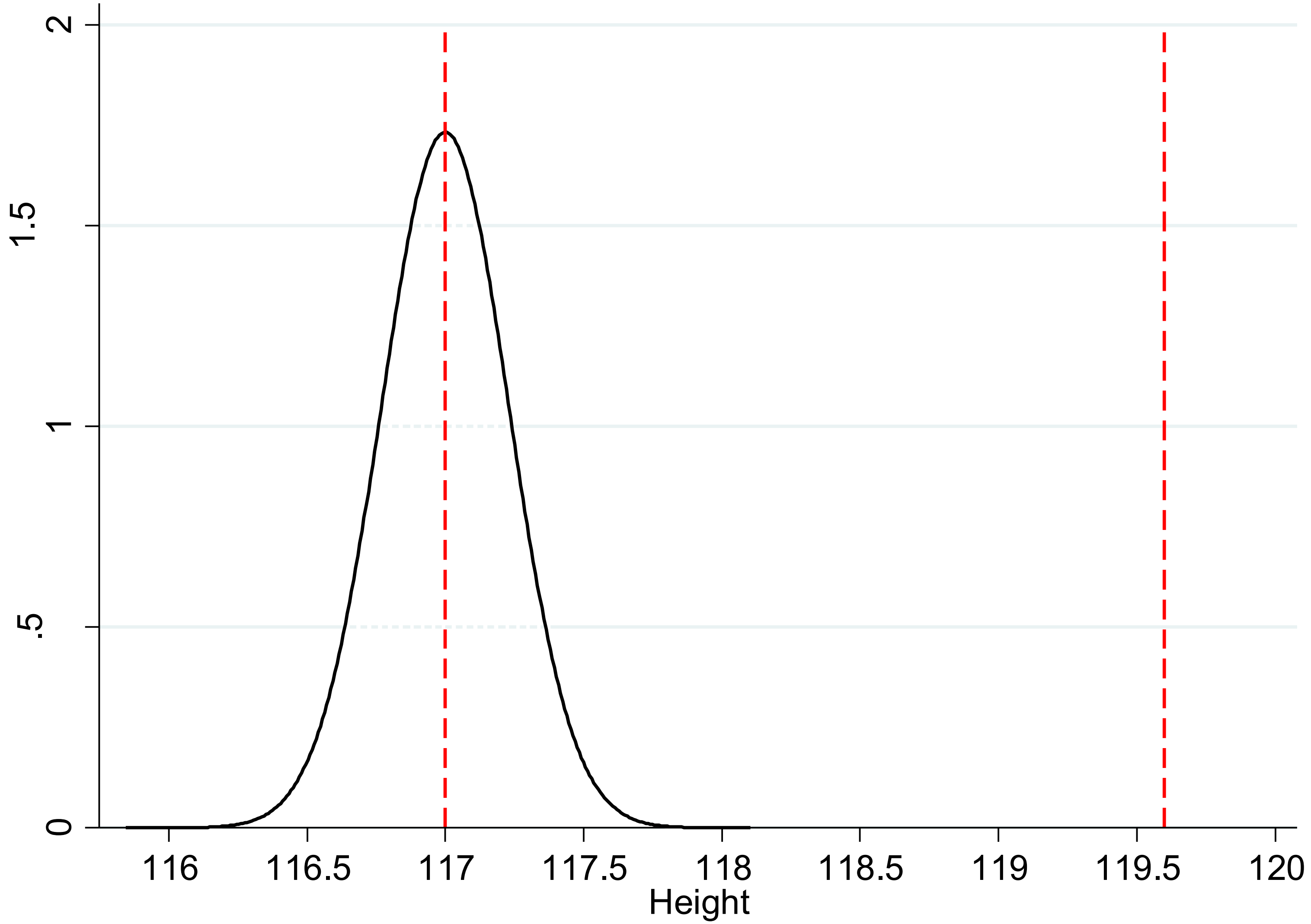
H1: μ≠117.0

- Στο δείγμα μας η μέση τιμή είναι 119.6
- Κάτω από τη μηδενική υπόθεση (μ=117.0), αν έπαιρνα πολλά δείγματα από τον πληθυσμό το 2015 η κατανομή των δειγματικών μέσων θα είχε μέση τιμή 117 και σταθερή απόκλιση το τυπικό σφάλμα (θα χρησιμοποιήσουμε το τυπικό σφάλμα που προκύπτει από το δικό μας δείγμα, δηλαδή SD/$\sqrt{n}$ =6.1/$\sqrt{730}$=0.23).
- Αφού έχουμε την κατανομή της μέσης τιμής, θα υπολογίσουμε πόσο πιθανό είναι να προκύψει η μέση τιμή που βρήκαμε στο δικό μας δείγμα μας (119.6 cm) αν ίσχυε η μηδενική υπόθεση

Δειγματοληπτική κατανομή των μέσων τιμών (μέση τιμή=117, SE=0.23)



Πόσο πιθανό είναι να προέκυψε η τιμή 119.6 cm το 2015 κατά τύχη (ενώ δηλαδή δεν μεταβλήθηκε το ύψος διαχρονικά);



Απίθανο!

One sample t-test για τη σύγκριση μέσης τιμής ενός δείγματος με θεωρητική (δοσμένη) τιμή

Το συγκεκριμένο είδος ερωτήματος (σύγκριση μέσης τιμής ενός δείγματος με θεωρητική τιμή) ελέγχεται με το **one sample t-test**

μ → μέση τιμή στον πληθυσμό (πχ. μέσο ανάστημα αγοριών 6 ετών το 2015)

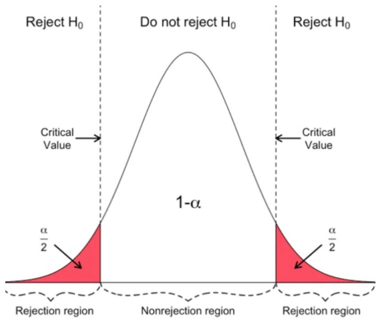
μ_0 → Μία δοσμένη τιμή (π.χ. το αντίστοιχο ύψος το 1981 → 117cm)

Μηδενική υπόθεση: $\mu = \mu_0$

Εναλλακτική υπόθεση: $\mu \neq \mu_0$

One sample t-test για τη σύγκριση μέσης τιμής ενός δείγματος με θεωρητική (δοσμένη) τιμή - διαδικασία

1. **Επιλέγουμε τυχαία ένα δείγμα αγοριών αυτής της ηλικίας** (n: μέγεθος δείγματος π.χ. N=730 αγόρια, $\mathbf{\bar{X}}$: μέση τιμή αναστήματος στο δείγμα π.χ. 119.60 cm)
2. **Η διαφορά από τη θεωρητική τιμή είναι** $|\bar{X} - \mu_0|$ (119.6-117.0=**2.6 cm**)
 Όσο πιο κοντά στο 0 η διαφορά → τόσο πιο πιθανό να μη διαφέρει το ανάστημα το 2015 σε σχέση με το 1981
3. Υ**πολογίζουμε την ποσότητα** $|\bar{X} - \mu_0|/SE$ (δηλαδή πόσα τυχαία σφάλματα βρίσκεται το 2.6 cm μακριά από το 0). Αν έπαιρνα πολλά τυχαία δείγματα 730 αγοριών το καθένα, η δειγματοληπτική κατανομή θα ήταν η κατανομή των $|\bar{X} - \mu_0|/SE$ που θα έβρισκα σε κάθε δείγμα. Κάτω από τη μηδενική υπόθεση, η μέση τιμή του $|\bar{X} - \mu_0|/SE$ θα είναι 0. **Η ποσότητα αυτή ακολουθεί την t κατανομή με n-1 βαθμούς ελευθερίας**
4. **Υπολογίζουμε την τιμή της t-κατανομής με n-1 βαθμούς ελευθερίας στο 5% επίπεδο σημαντικότητας (→ κρίσιμη τιμή)**
5. **Αν η πιθανότητα να προκύψει από τα δεδομένα μας η τιμή** $|\bar{X} - \mu_0|/SE$ **είναι ≤5%** → στατιστικά σημαντική διαφορά (απορρίπτω τη μηδενική υπόθεση)

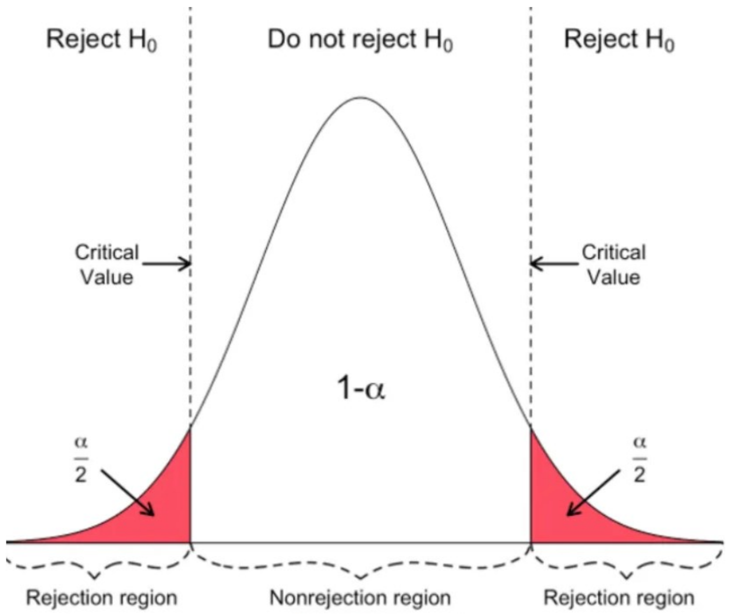


One sample t-test: Βήματα για την υλοποίηση

1. Υπολογίζουμε την ποσότητα $|\bar{X} - \mu_0|/SE$ από το δείγμα μας
2. Βρίσκουμε από πίνακες τη θεωρητική τιμή $t_{n-1,0.05}$
3. Συγκρίνουμε:

- **Αν $|\bar{X} - \mu_0|/SE$ >** $t_{n-1,0.05}$
  → η διαφορά είναι στατιστικά σημαντική
  (απορρίπτω τη μηδενική υπόθεση)

- **Αν $|\bar{X} - \mu_0|/SE$ <** $t_{n-1,0.05}$
  → η διαφορά δεν είναι στατιστικά σημαντική
  →μπορεί δηλαδή να οφείλεται στην τύχη
  (δεν απορρίπτω τη μηδενική υπόθεση)



Η συνάρτηση $|\bar{X} - \mu_0|/SE$ αποτελεί τη συνάρτηση του στατιστικού κριτηρίου και συγκεκριμένα **το στατιστικό κριτήριο t.**

Προϋπόθεση για την εφαρμογή του one sample t-test

- Το ποσοτικό χαρακτηριστικό που μελετάμε να ακολουθεί την κανονική κατανομή

Στο παράδειγμά μας

- 1981 → μέσο ύψος των αγοριών ηλικίας 6 ετών: 117.0 cm
- Το 2015, δείγμα n=730 αγόρια, Μέση τιμή (SD): 119.6 (6.1) cm

H0: μ=117.0

H1: μ≠117.0

1. Υπολογισμός t= $|\bar{X} - \mu_0|$/SE =
   =|119.6-117|/(6.1/$\sqrt{730}$)=**11.3**
2. Βρίσκω την τιμή της t κατανομής για **n-1**=730-1=**729** βαθμούς ελευθερίας και για α=5% από τους αντίστοιχους πίνακες της t κατανομής → 1.96
3. Διαπιστώνω ότι $|\bar{X} - \mu_0|$/SE =**11.3>1.96**
4. **Απορρίπτω** τη μηδενική υπόθεση → άρα το 2015 τα παιδιά πιο ψηλά από ότι το 1981

Degrees of freedom	Significance level 10%	5%	1%
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.894	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
......	......	......	......
60	1.671	2.000	2.660
70	1.667	1.994	2.648
80	1.664	1.990	2.639
90	1.662	1.987	2.632
100	1.660	1.984	2.626
120	1.658	1.980	2.617
150	1.655	1.976	2.609
200	1.653	1.972	2.601
300	1.650	1.968	2.592
400	1.649	1.966	2.588
500	1.648	1.965	2.586
600	1.647	1.964	2.584
∞	1.645	1.960	2.576