Τμήμα Διεθνών & Ευρωπαϊκών Σπουδών
Πανεπιστήμιο Πειραιώς

Σημειώσεις Ποσοτικών Μεθόδων

Ιωάννης Αθ. Παραβάντης
Καθηγητής

Εαρινό εξάμηνο 2023-2024

Διδάσκων

Επικοινωνία με διδάσκοντα

- Γιάννης Παραβάντης
  - *"Κύριε Παραβάντη"* ή *"Δάσκαλε"*
- Πολιτικός Μηχανικός – Συγκοινωνιολόγος (ΕΜΠ), MSc Transportation & PhD Environmental Health (Northwestern University)
- Βαθμίδα: Καθηγητής (στο ΠΑΠΕΙ από το 2020)
- Γνωστικό αντικείμενο: *"Ενέργεια, Τεχνολογία, και Περιβάλλον στη Διεθνή Πολιτική"*
- Γραφείο 105, 1ος όροφος, Οδυσσέα Ανδρούτσου 150
- Τηλ. 210 414-2771 & 697 3048824 (και για Viber)
- Emails: paravantis@gmail.com & jparav@unipi.gr
- Ώρες γραφείου θα ανακοινωθούν

Κριτικές παλαιότερων φοιτητών

"Αυστηρός + ΠΡΟΘΥΜΟΣ! Το μυστικό είναι να μη τον ντρέπεσαι και να τον ρωτάς απορίες. Αν η απορία είναι low level, θα σε παραπέμψει ευγενικά. Αν όχι, θα είναι θησαυρός. Αν δεν τον ρωτάς, θα πελαγώσεις. Αν τον κοροϊδέψεις, την πάτησες."

Κριτική από το www.rateyourprofessor.gr
(το site δεν υπάρχει πια)

"... κύριε Παραβάντη σας στέλνω για να σας πω ότι πλέον φοιτώ στην ΑΣΟΕΕ στο μεταπτυχιακό Διοίκηση Ανθρώπινου Δυναμικού για το οποίο μου γράψατε συστατική επιστολή και σας ευχαριστώ! Αρχίσαμε Ποσοτική Μεθοδολογίας Έρευνας και έχω να πω ότι είμαι τυχερή που παρακολουθούσα γιατί ακόμα θυμάμαι τις παραδόσεις σας. Σας ευχαριστώ."

Μήνυμα αποφοίτου

"Μια συμβουλή: Τα μαθήματα του Παραβάντη είναι ευαγγέλιο. Γίνε αστέρι σε αυτά. Είναι πάρα πολύ χρήσιμα στο εξωτερικό."

Δήλωση φοιτητή Erasmus στο Πανεπιστήμιο του Plymouth (4ο έτος ΔΕΣ)

---

1. Μάθημα

1.1. Συγγράμματα

Οικονόμου, Γ. & Αγιακλόγλου, Χ.Ν. (2004). *Μέθοδοι Προβλέψεων και Ανάλυσης Αποφάσεων*. Εκδόσεις Μπένου

- Πιο μαθηματικό, πιο συνοπτικό

Aczel, A. & Sounderpandian, J. (2013). *Στατιστική σκέψη στον κόσμο των επιχειρήσεων*. Μ. Σφακιανάκης, Επιμέλεια-Πρόλογος Ελληνικής Έκδοσης, Broken Hill Publishers & Εκδόσεις Πασχαλίδης

- Πιο αναλυτικό, πολλές λυμένες ασκήσεις

1.2. Διδακτικό υλικό

Ανακοινώσεις, διαφάνειες, δεδομένα, φύλλα υπολογισμοί, κλπ. θα τα διαβάζετε και κατεβάζετε όλα από το eclass του μαθήματος

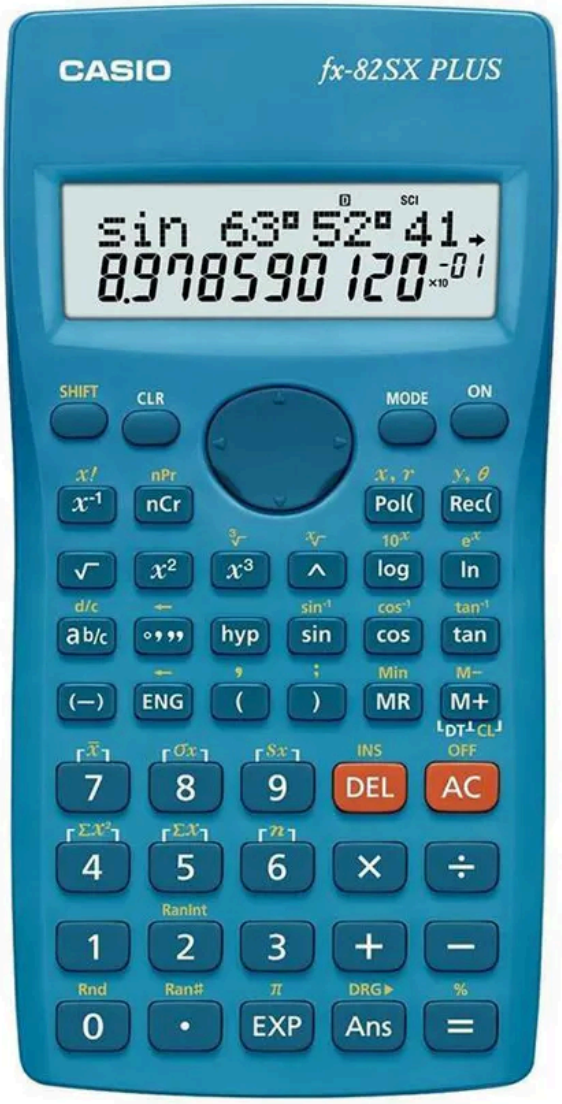
- https://eclass.unipi.gr/courses/DES103/

1.3. Αξιολόγηση

Μορφές αξιολόγησης

- Πρόοδοι (μια έως τρεις, όλες multiple choice, με προειδοποίηση στο προηγούμενο μάθημα) με ανοιχτό όλο το εκπαιδευτικό υλικό και εξ αποστάσεως (take home): **30%** του τελικού βαθμού
    - *Για να μπορέσετε να συμμετάσχετε στις προόδους, θα πρέπει να έχετε και να κάνετε χρήση των credentials (username, password) που σας έχει δώσει το Πανεπιστήμιο Πειραιώς (και με τα οποία μπαίνετε στο φάκελό σας και βλέπετε τους βαθμούς σας)*
- Τελική εξέταση (multiple choice με ασκήσεις) με ανοιχτό όλο το εκπαιδευτικό υλικό και δια ζώσης: **70%** του τελικού βαθμού
- Bonus παρουσίας στην τάξη και συμμετοχής στο μάθημα (κατά την κρίση του διδάσκοντα): **10%** του τελικού βαθμού
    - Να είστε παρόντες στις διαλέξεις και να συμμετέχετε με ερωτήσεις και απαντήσεις

Στην (δια ζώσης) τελική εξέταση θα πρέπει να έχετε μαζί σας ένα (φθηνό) επιστημονικό κομπιουτεράκι (calculator)



1.4. Λογισμικό που θα χρησιμοποιηθεί

Google Sheets

- Αν δεν έχετε λογαριασμό στο Gmail, παρακαλώ να φτιάξετε (είναι δωρεάν)
- Ιδίως θα χρησιμοποιήσουμε ένα φύλλο υπολογισμού Google Sheets, που το έχω ονομάσει DESCSTAT (DESCriptive STATistics) και βρίσκεται εδώ
    - Το φύλλο αυτό μπορείτε να το αντιγράψετε και στο δικό σας Google Drive ή να το κατεβάσετε ως Excel

---

1.5. Δεκαδικό ψηφίο και χωρισμός χιλιάδων

Στις σημειώσεις αυτές ακολουθείται το σύστημα των ΗΠΑ

- Για να χωρίζουμε το ακέραιο μέρος ενός αριθμού από τα δεκαδικά χρησιμοποιείται η τελεία “.” (και όχι η κόμμα, όπως συνηθίζεται στην Ελλάδα)
- Για των χωρισμό των χιλιάδων σε ένα μεγάλο αριθμό χρησιμοποιείται το κόμμα (“,”) και όχι η τελεία, όπως συνηθίζεται στην Ελλάδα
- Χωρίζουμε με κόμμα τις χιλιάδες ακόμα και σε αριθμούς που έχουν ακέραιο μέρος με 4 μόνο ψηφία
    - (Μερικά εγχειρίδια μορφοποίησης συστήνουν να μην χωρίζουμε τις χιλιάδες για αριθμούς που έχουν 4 μόνο ψηφία)

Παραδείγματα

- Ο αριθμός χίλια διακόσια τριάντα τέσσερα και πενήντα έξι γράφεται ως 1,234.56
    - (Κάποια στυλ μορφοποίησης θα συνιστούσαν, επειδή το ακέραιο μέρος, “1234”, δεν είναι πάνω από τέσσερα ψηφία, να μην χωρίσουμε τις χιλιάδες)
- Ο αριθμός δώδεκα χιλιάδες τριακόσια σαράντα πέντε και εξήντα επτά γράφεται ως 12,345.67
- Το σύστημα αυτό δεν διαφέρει πολύ από το σύστημα που συνιστά η Ευρωπαϊκή Ένωση, δηλαδή να χωρίζονται οι χιλιάδες με κενό (space) αντί για κόμμα, και να γίνεται χρήση είτε κόμμα είτε τελείας για να χωρίζει τα δεκαδικά

1.6. Εάν δεν τα πάτε καλά με τα μαθηματικά

Ούτε ο διδάσκων τα πάει καλά με τα μαθηματικά

Το μάθημα διδάσκεται ως **Ανάλυση Δεδομένων** (data analysis) όχι **Μαθηματική Στατιστική** (Mathematical Statistics)

- Το μάθημα το διδάσκει μηχανικός, όχι μαθηματικός

Δεν θα υπάρξει παρά ελάχιστη κάλυψη πιθανοτήτων (probability), γεγονός που αφαιρεί ένα σημαντικό βαθμό δυσκολίας

Θα δοθεί έμφαση στη χρήση δωρεάν υπολογιστικών εργαλείων (Google Sheets)

Σημαντικό να γίνεται το μάθημα κατανοητό κατά τη διάρκεια της παράδοσης

- Αλλά όταν μπούμε στην **Επαγωγική Στατιστική**, εάν παράλληλα με την παρακολούθηση δεν διαβάζετε, θα χάσετε τη μπάλα
- Θα ξέρετε πότε θα έρθει αυτή η στιγμή γιατί θα υπάρξει πρόοδος με το τέλος της **Περιγραφικής Στατιστικής**

Θα βοηθήσει πολύ να κάνετε χρήση υπολογιστή καθώς διαβάζετε

- Statistics is learned by doing

1.7. Πως θα περάσετε το μάθημα

Να έρχεστε ανελλιπώς στις παραδόσεις

- Περνάμε καλά, δεν θα το βλέπετε σαν αγγαρεία
- Εξάλλου το ΠΑΠΕΙ είναι η πρώτη σας δουλειά

Να διαβάσετε για την πρόοδο (ή τις προόδους)

- Για να εξασφαλίσετε καλό βαθμό
- Για να μην μείνουν όλα στο τέλος και πρέπει να μάθετε ένα ολόκληρο βιβλίο σε τρεις μέρες
    - Θα είναι σαν να πρέπει να πιείτε ένα ολόκληρο βαρέλι σε δύο μέρες

Να επικοινωνείτε με τον διδάσκοντα

- Να έρχεστε με απορίες στις ώρες γραφείου
    - Μπορούμε να συναντιόμαστε και online (Google Meet)
- Να στέλνετε email με απορίες στον καθηγητή

---

2. Βασικές έννοιες & είδη στατιστικής ανάλυσης

Ποσοτικές Μέθοδοι (Quantitative Methods) = Στατιστική (Statistics) = Ανάλυση δεδομένων (Data analysis)

- Η τεχνητή νοημοσύνη (Artificial Intelligence ή AI) βασίζεται σε μεγάλο μέρος στη στατιστική και στις πιθανότητες

Στις Ποσοτικές Μεθόδους αναλύσουμε **δεδομένα** (data)

- Τα δεδομένα είναι οργανωμένα σε **παρατηρήσεις** (observations ή cases), μια για κάθε στατιστική μονάδα (ή στατιστικό άτομο)

3 variables

15 observations

Weight (pounds)	Length (inches)	Region
290	30	East
296	35	East
299	34	East
300	34	East
305	38	East
307	40	North
311	46	North
315	45	North
325	49	North
339	48	North
340	55	South
355	58	South
357	55	West
359	57	West
361	59	West

	Weight (pounds)	Length (inches)	Region
	290	30	East
Observation #2	296	35	East
	299	34	East
	300	34	East
	305	38	East
	307	40	North
	311	46	North
	315	45	North
	325	49	North
	339	48	North
	340	55	South
	355	58	South
	357	55	West
	359	57	West
	361	59	West

Είδη στατιστικής:

1. **Περιγραφική** (descriptive)
2. **Επαγωγική** (inferential)

Με ενδιαφέρει περισσότερο να μάθετε τους Αγγλικούς παρά τους Ελληνικούς όρους

2.1. Περιγραφική στατιστική

Ενδιαφερόμαστε μόνο για ένα **δείγμα** (sample), που είναι ένα σύνολο δεδομένων (data set)

Δεν μας ενδιαφέρει ο **πληθυσμός** (population) από τον οποίο προέρχεται το δείγμα

Αναλύουμε **μεταβλητές** (variables, μια ή περισσότερες)

Τα δεδομένα και οι μεταβλητές είναι οργανωμένες σε **παρατηρήσεις** (observations ή cases)

Γραφικές μέθοδοι

- Boxplot, ιστόγραμμα (histogram), ραβδόγραμμα (barchart)

Monthly Deaths from Lung Diseases in the UK

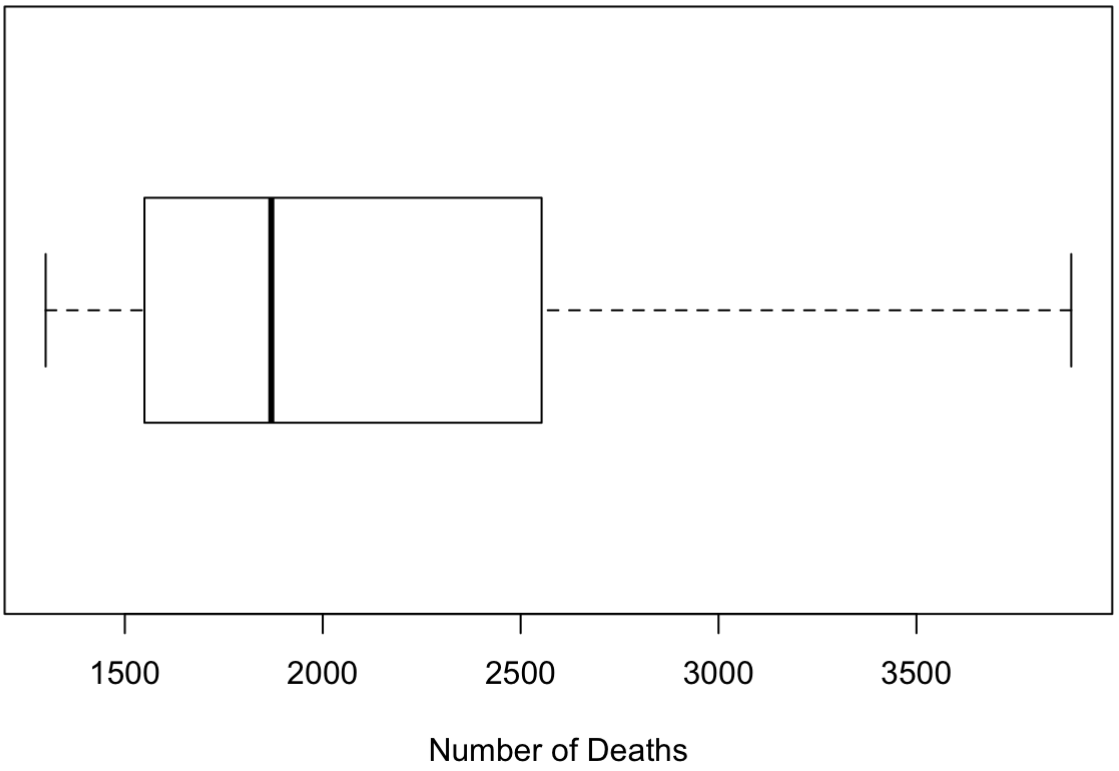
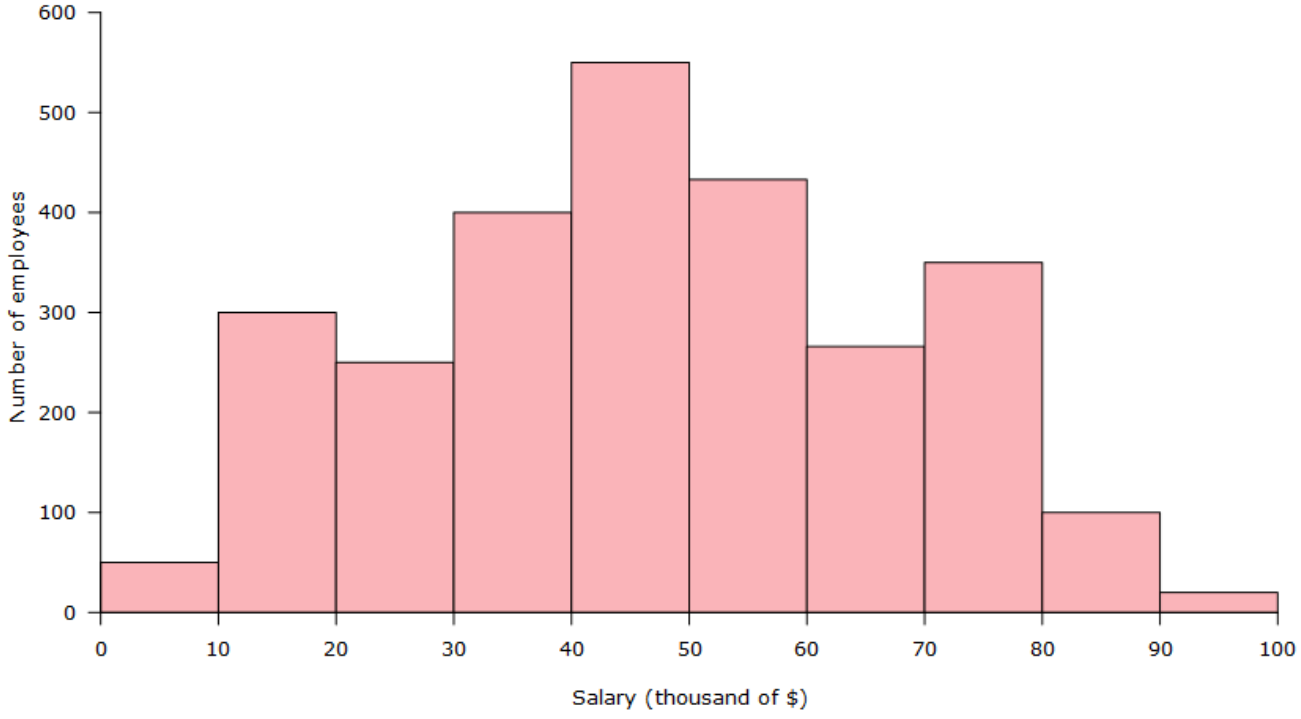
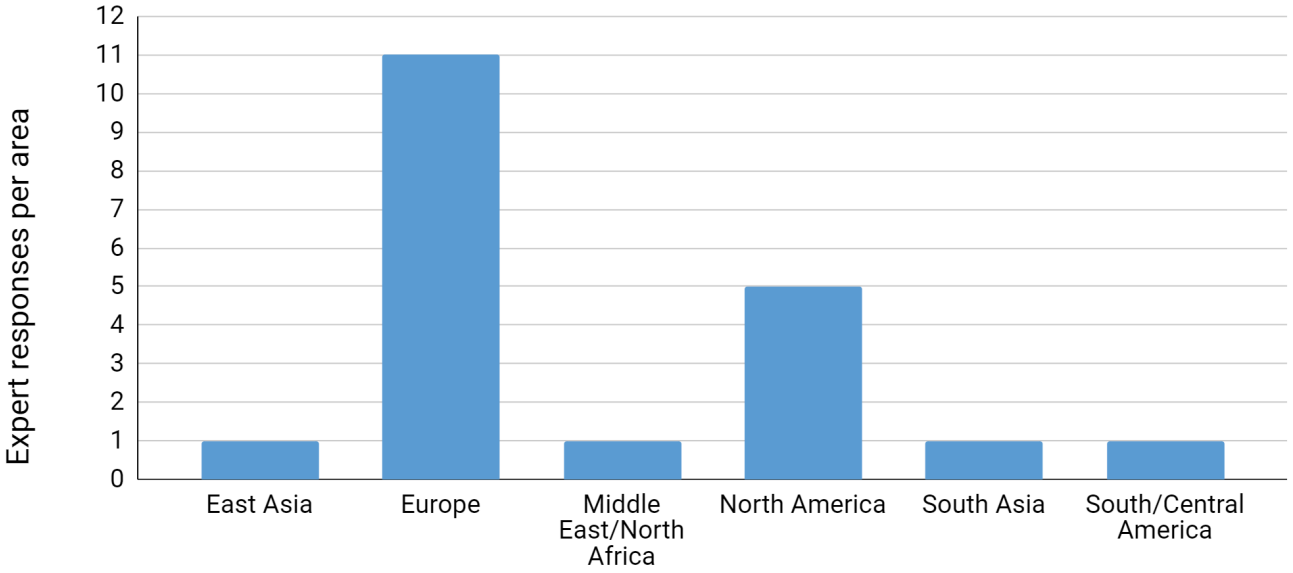


Chart 5.7.1
Distribution of salaries of the employees of ABC Corporation





Αριθμητικές μέθοδοι

- Μέτρα θέσης και διασποράς
  - Π.χ. ελάχιστο (minimum), μέγιστο (maximum), μέση τιμή (mean ή average), τυπική απόκλιση (standard deviation)

Θα δούμε και κάποιους πίνακες συχνοτήτων (frequency tables)

Cumulative Frequency Distribution Table

cuemath
THE MATH EXPERT

Class Interval	Tally Marks	Frequency	Cumulative Frequency (cf)
20 - 30	||	2	2
30 - 40	||||	4	6
40 - 50	||||		6	12
50 - 60	||	2	14
60 - 70	||||		6	20
70 - 80	||||	4	24

2.2. Επαγωγική στατιστική

Εξετάζουμε ένα **δείγμα** (sample) για να καταλάβουμε τον **πληθυσμό** (population)

Αναλύουμε (συνήθως) περισσότερες από μια **μεταβλητές** (variables)

Βασιζόμαστε στις γραφικές μεθόδους της περιγραφικής στατιστικής

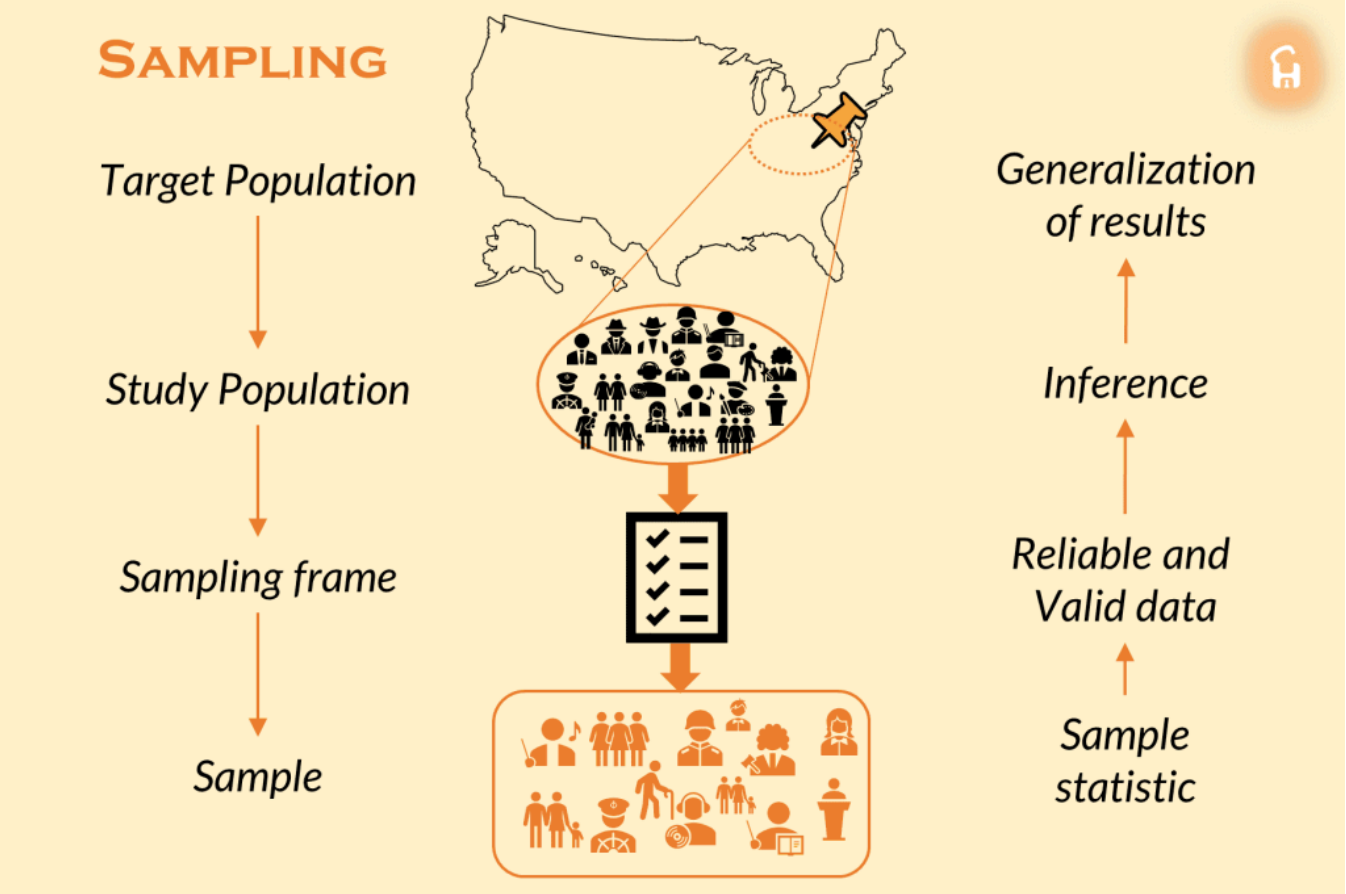
- Κυρίως στο ιστόγραμμα (histogram)

Βασιζόμαστε επίσης σε αριθμητικά μεγέθη που υπολογίζουμε στην περιγραφική στατιστική

- Κυρίως **μέση τιμή** (mean ή average) και **τυπική απόκλιση** (standard deviation)

Διενεργούμε στατιστικούς ελέγχους (statistical tests)

- Μέσω αυτών, πάμε από δειγματικά μεγέθη (sample statistics) σε πληθυσμιακές παραμέτρους (population parameters)



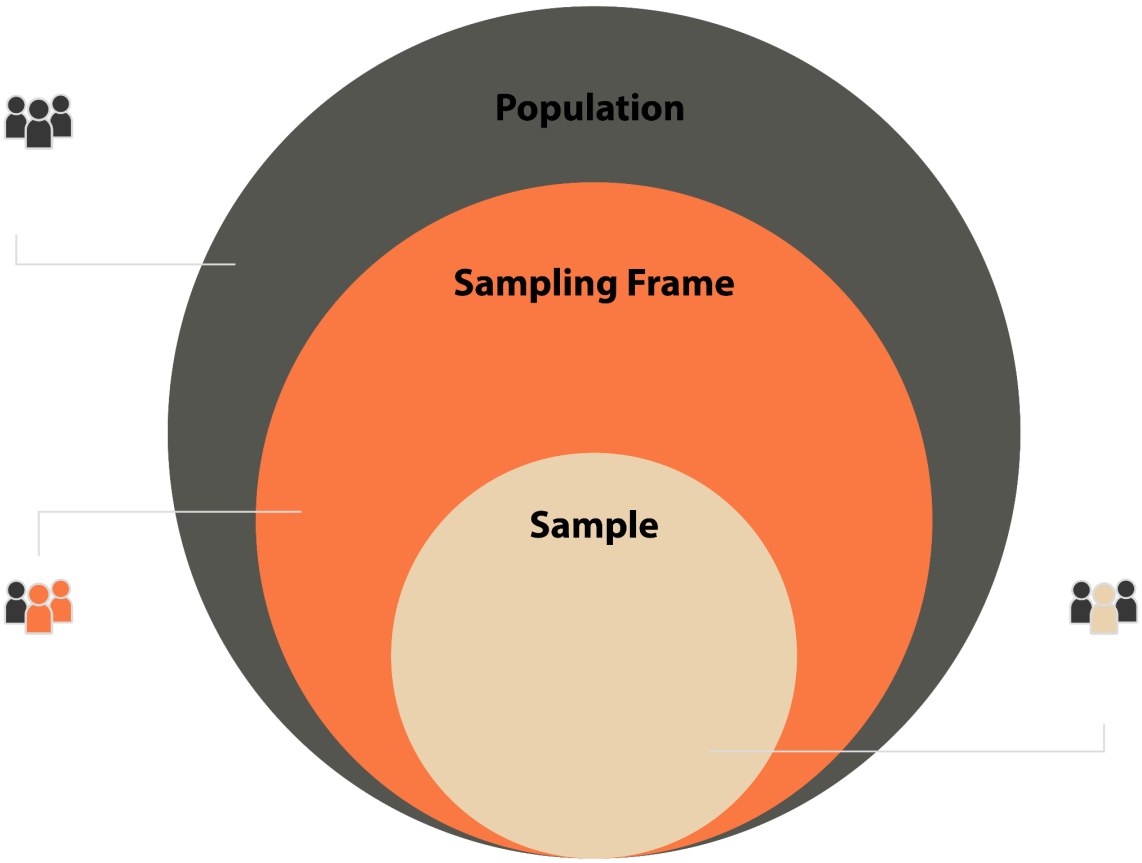
---

3. Πληθυσμός και δείγμα

Πληθυσμός (population)

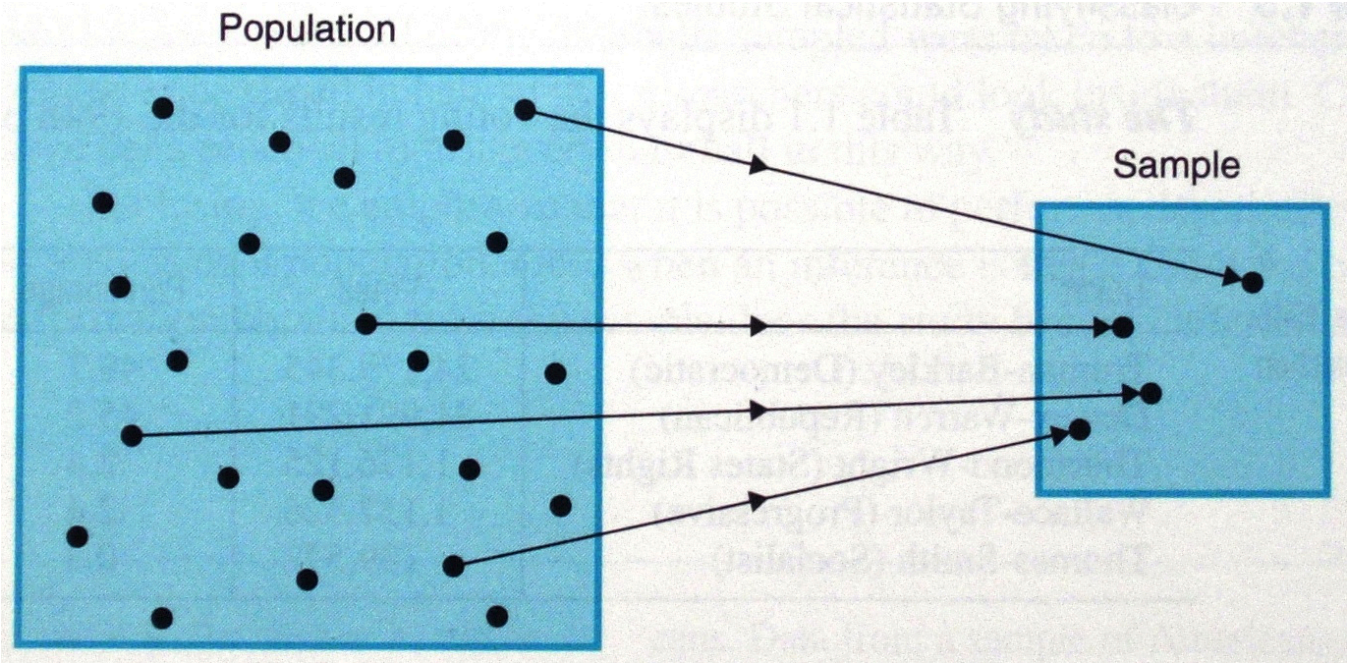
Δειγματοληπτικό πλαίσιο (sampling frame)

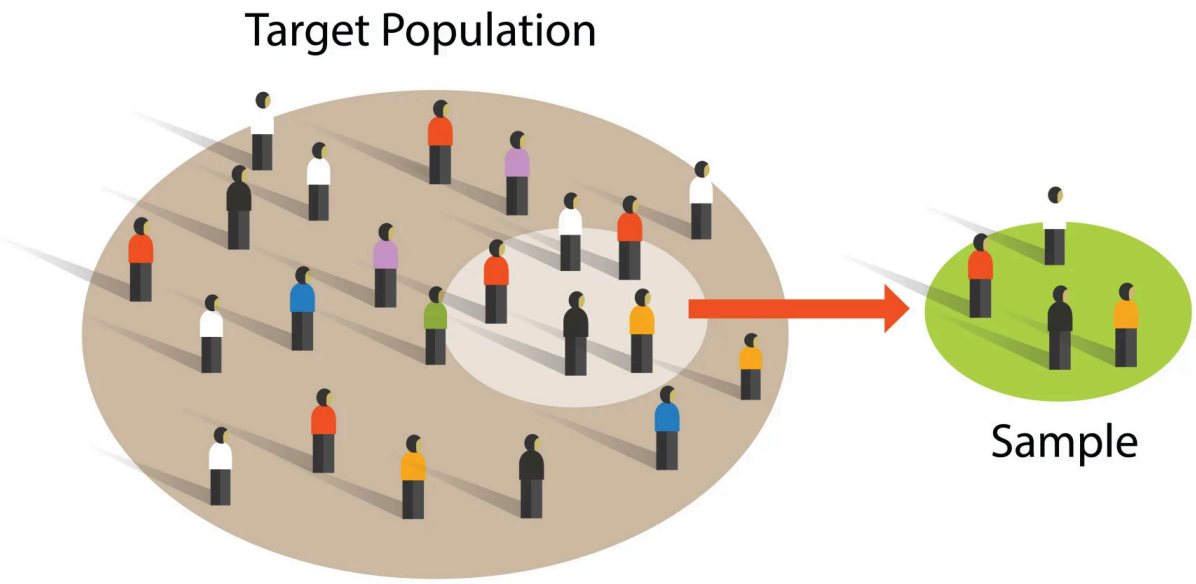
- Το τμήμα εκείνο του πληθυσμού στο οποίο έχουμε πρόσβαση (π.χ. Αν πρόκειται για άτομα, διαθέτουν κινητό τηλέφωνο) και από το οποίο μπορούμε να επιλέξουμε δεδομένα για το δείγμα

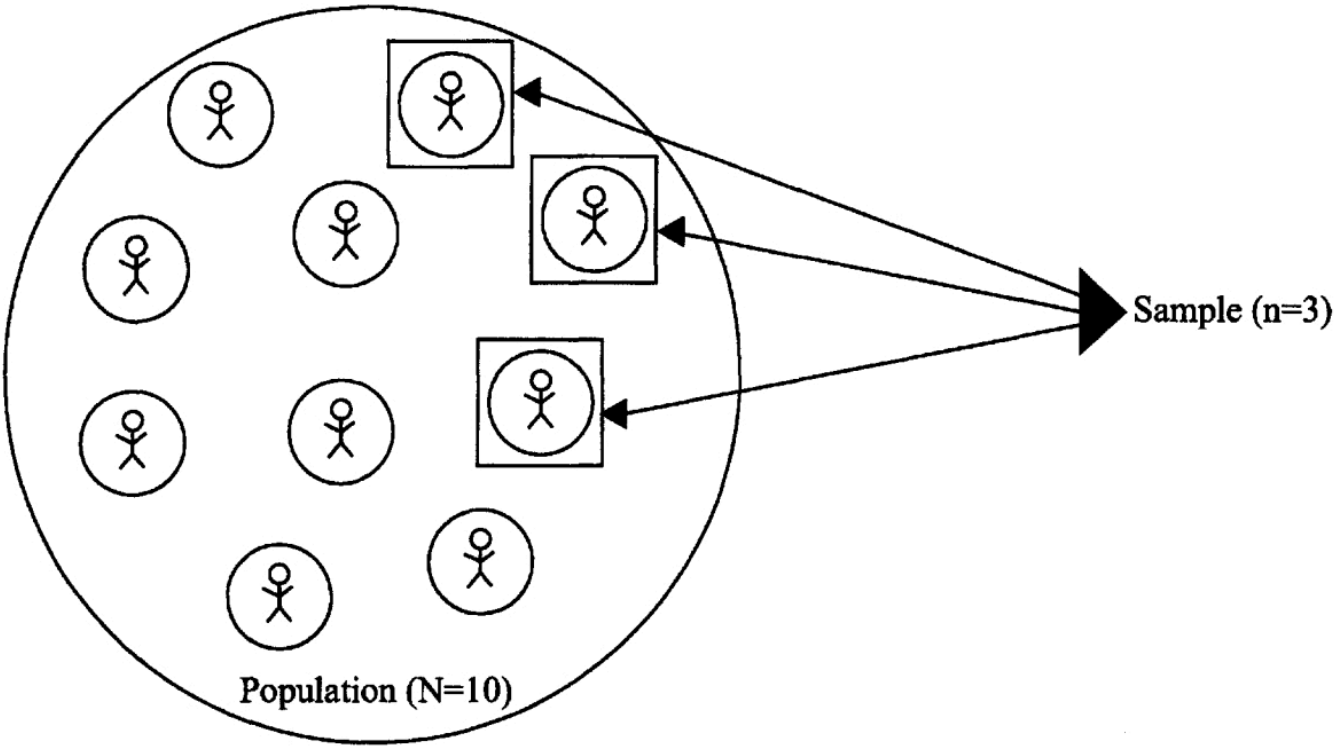


Όταν κάνουμε **απογραφή** (census) του πληθυσμού, καταγράφουμε όλες τις στατιστικές μονάδες που περιλαμβάνονται στον πληθυσμό (στην πραγματικότητα στο δειγματοληπτικό πλαίσιο, κάποιες στατιστικές μονάδες πάντα ξεφεύγουν)

Ένα **δείγμα** (sample) είναι πάντα πολύ μικρότερο (“τάξεις μεγέθους”) του πληθυσμού







Παρά το γεγονός ότι στις παραπάνω εικόνες φαίνεται να επιλέγεται δείγμα από τον πληθυσμό, να θυμάστε ότι το δείγμα επιλέγεται πάντα από το δειγματοληπτικό πλαίσιο με την ελπίδα:

- Το τμήμα του πληθυσμού που βρίσκεται εκτός δειγματοληπτικού πλαισίου να είναι όσο πιο μικρό γίνεται
- Οι στατιστικές μονάδες που βρίσκονται εκτός δειγματοληπτικού πλαισίου να μην έχουν συστηματικές διαφορές από τις στατιστικές μονάδες που βρίσκονται εντός δειγματοληπτικού πλαισίου!

3.1. Τυχαίο δείγμα

Ένα δείγμα πρέπει να συλλέγεται κατά τέτοιο τρόπο ώστε να είναι **αντιπροσωπευτικό** (representative), δηλαδή τα χαρακτηριστικά του να μοιάζουν με αυτά του πληθυσμού

Το πιο κλασικό είδος **πιθανοτικής δειγματοληψίας** (probabilistic sampling) είναι το **απλό τυχαίο δείγμα** (simple random sample ή SRS), που έχει τα ακόλουθα χαρακτηριστικά:

- Κάθε στατιστική μονάδα (ή άτομο) έχει την ίδια πιθανότητα επιλογής
- Κάθε δείγμα του ιδίου μεγέθους έχει την ίδια πιθανότητα επιλογής

Για την επιλογή απλού τυχαίου δείγματος χρησιμοποιούνται **τυχαίοι αριθμοί** (random numbers)

TABLE 1.9
Random numbers

Line number	Column number									
	00–09		10–19		20–29		30–39		40–49	
00	15544	80712	97742	21500	97081	42451	50623	56071	28882	28739
01	01011	21285	04729	39986	73150	31548	30168	76189	56996	19210
02	47435	53308	40718	29050	74858	64517	93573	51058	68501	42723
03	91312	75137	86274	59834	69844	19853	06917	17413	44474	86530
04	12775	08768	80791	16298	22934	09630	98862	39746	64623	32768
05	31466	43761	94872	92230	52367	13205	38634	55882	77518	36252
06	09300	43847	40881	51243	97810	18903	53914	31688	06220	40422
07	73582	13810	57784	72454	68997	72229	30340	08844	53924	89630
08	11092	81392	58189	22697	41063	09451	09789	00637	06450	85990
09	93322	98567	00116	35605	66790	52965	62877	21740	56476	49296
10	80134	12484	67089	08674	70753	90959	45842	59844	45214	36505
11	97888	31797	95037	84400	76041	96668	75920	68482	56855	97417
12	92612	27082	59459	69380	98654	20407	88151	56263	27126	63797
13	72744	45586	43279	44218	83638	05422	00995	70217	78925	39097
14	96256	70653	45285	26293	78305	80252	03625	40159	68760	84716
15	07851	47452	66742	83331	54701	06573	98169	37499	67756	68301
16	25594	41552	96475	56151	02089	33748	65289	89956	89559	33687
17	65358	15155	59374	80940	03411	94656	69440	47156	77115	99463
18	09402	31008	53424	21928	02198	61201	02457	87214	59750	51330
19	97424	90765	01634	37328	41243	33564	17884	94747	93650	77668

Σπανίως γίνεται χρήση απλών τυχαίων δειγμάτων στην πράξη

Κατά κανόνα γίνεται χρήση **τυχαίου δείγματος** (random sample), που έχει τις εξής ιδιότητες:

- Κάθε στατιστική μονάδα (ή άτομο) έχει την ίδια πιθανότητα επιλογής (όπως και στο απλό τυχαίο δείγμα)
- Όμως, δεν έχει κάθε δείγμα του ιδίου μεγέθους την ίδια πιθανότητα επιλογής!

Αυτό ακούγεται μπερδευτικό, οπότε ένα παράδειγμα βοηθάει να διακρίνουμε αυτές τις δύο έννοιες:

"You have a class of 50 students, 30 males and 20 females. You are randomly selecting 3 males and 2 females. This is an example of a ***random sample*** *because each person in the class has the same probability of being selected 3/30=2/20. However, all your samples are going to be the same: 3 males and 2 females. It is not possible to get 2 males and 3 females, or 4 males and 1 female. Therefore, it is not a* ***simple random sample****."*

---

4. Δειγματοληψία

Η διαδικασία επιλογής ενός δείγματος ονομάζεται **δειγματοληψία** (sampling)

Η δειγματοληψία έχει δύο στόχους:

1. Το δείγμα που θα επιλεγεί να είναι όσο γίνεται πιο αντιπροσωπευτικό του πληθυσμού
2. Η δειγματοληψία να έχει όσο γίνεται μικρότερο κόστος

Όπως είπαμε προηγουμένως, σχεδόν ποτέ δεν επιλέγεται απλό τυχαίο δείγμα

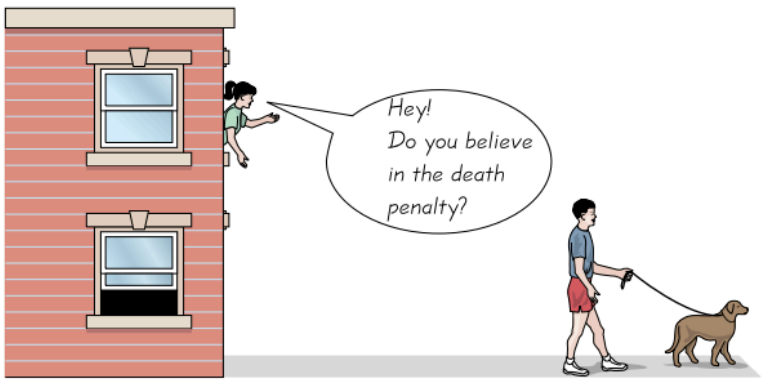
- Η επιλογή απλού τυχαίου δείγματος δεν θα διασφάλιζε κανέναν από τους δύο προηγούμενους στόχους

Επειδή σχεδόν ποτέ δεν επιλέγεται απλό τυχαίο δείγμα, αναφερόμαστε στις διάφορες μορφές δειγματοληψία και σαν **ψευδοτυχαία δειγματοληψία** (pseudorandom sampling)

Παρουσιάζονται στις επόμενες ενότητες τα πιο συνηθισμένα είδη δειγματοληψίας

4.1. Δειγματοληψία ευκολίας

Στη **δειγματοληψία ευκολίας** (convenience sampling) επιλέγεται δείγμα που βολεύει τους ερευνητές



Convenience Sampling:
Use results that are easy to get.

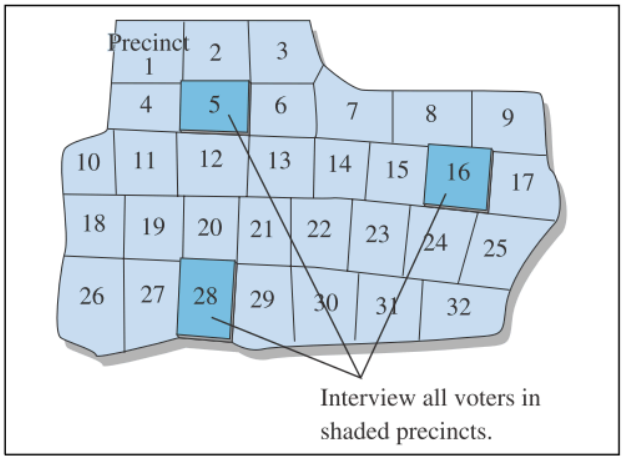
4.2. Συστηματική δειγματοληψία

Στη **συστηματική δειγματοληψία** (systematic sampling), π.χ. ανοίγουμε τον τηλεφωνικό κατάλογο και επιλέγουμε κάθε εικοστό όνομα

Systematic Sampling:
Select some starting point, then select every kth (such as every 50th) element in the population.

4.3. Δειγματοληψία κατά στιβάδες

Στη **δειγματοληψία κατά στιβάδες** (cluster sampling), ο πληθυσμός χωρίζεται σε στιβάδες (π.χ. γεωγραφικές περιοχές) και μετά επιλέγονται όλα τα μέλη κάθε στοιβάδας



Cluster Sampling:
Divide the population area into sections (or clusters), then randomly select some of those clusters, and then choose all members from those selected clusters.

Αν οι στιβάδες είναι μεγάλες, τότε επιλέγεται τυχαίο δείγμα από κάθε στιβάδα, οπότε μιλάμε για **πολυσταδιακή δειγματοληψία** (multistage sampling), που παρουσιάζεται παρακάτω

4.4. Στρωματοποιημένη δειγματοληψία

Τέλος, στη **στρωματοποιημένη δειγματοληψία** (stratified sampling), ο πληθυσμός χωρίζεται σε (ομοειδείς) ομάδες, π.χ. άνδρες και γυναίκες, ακολούθως δε επιλέγεται τυχαίο δείγμα από κάθε υποομάδα

Stratified Sampling:
Subdivide the population into at least two different subgroups (or strata) that share the same characteristics (such as gender or age bracket), then draw a sample from each subgroup.

Σχετικό παράδειγμα κειμένου από παραδοτέο (deliverable) της ερευνητικής ομάδας μου στο πλαίσιο ερευνητικού έργου Horizon 2020, που σχολιάζει μια βιβλιογραφική πηγή:

Using a qualitative approach based on 37 in-depth semi-structured interviews conducted in 2019 and 2020, the study focused on three themes: energy supply security, the renewables' role in enhancing energy security, and the connections between energy security and environmental sustainability. Approximately one-third of the participants were purposefully selected to ensure diversity in their backgrounds and professions, with a particular focus on academics. The remaining two-thirds were chosen through random sampling to ensure an unbiased representation of the broader population. The two main criteria for including a participant in the sample were their willingness to discuss energy security and share personal views on the topic and a completed undergraduate university degree.

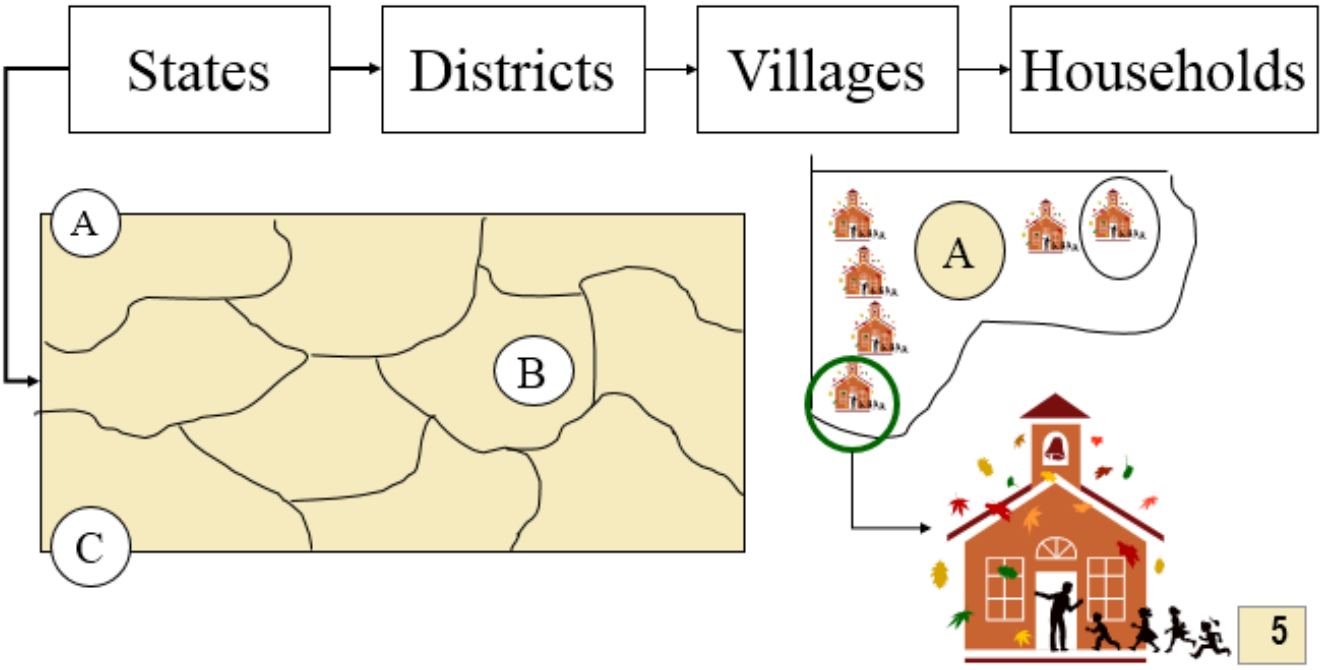
Η πηγή είναι:

- Mouraviev, N. (2021). Energy security in Kazakhstan: The consumers' perspective. *Energy Policy*, 155, 112343. https://doi.org/10.1016/j.enpol.2021.112343

Η μορφοποίηση (formatting) της παραπάνω πηγής έχει γίνει σύμφωνα με το στυλ APA (American Psychological Association), που είναι ένα από τα επικρατέστερα πρότυπα στις κοινωνικές επιστήμες.

4.5. Πολυσταδιακή δειγματοληψία

Συχνά, οι παραπάνω μορφές δειγματοληψίας συνδυάζονται μεταξύ τους περιλαμβάνοντας και κάποια τυχαία δείγματα, οπότε μιλάμε για **πολυσταδιακή δειγματοληψία** (multistage sampling)



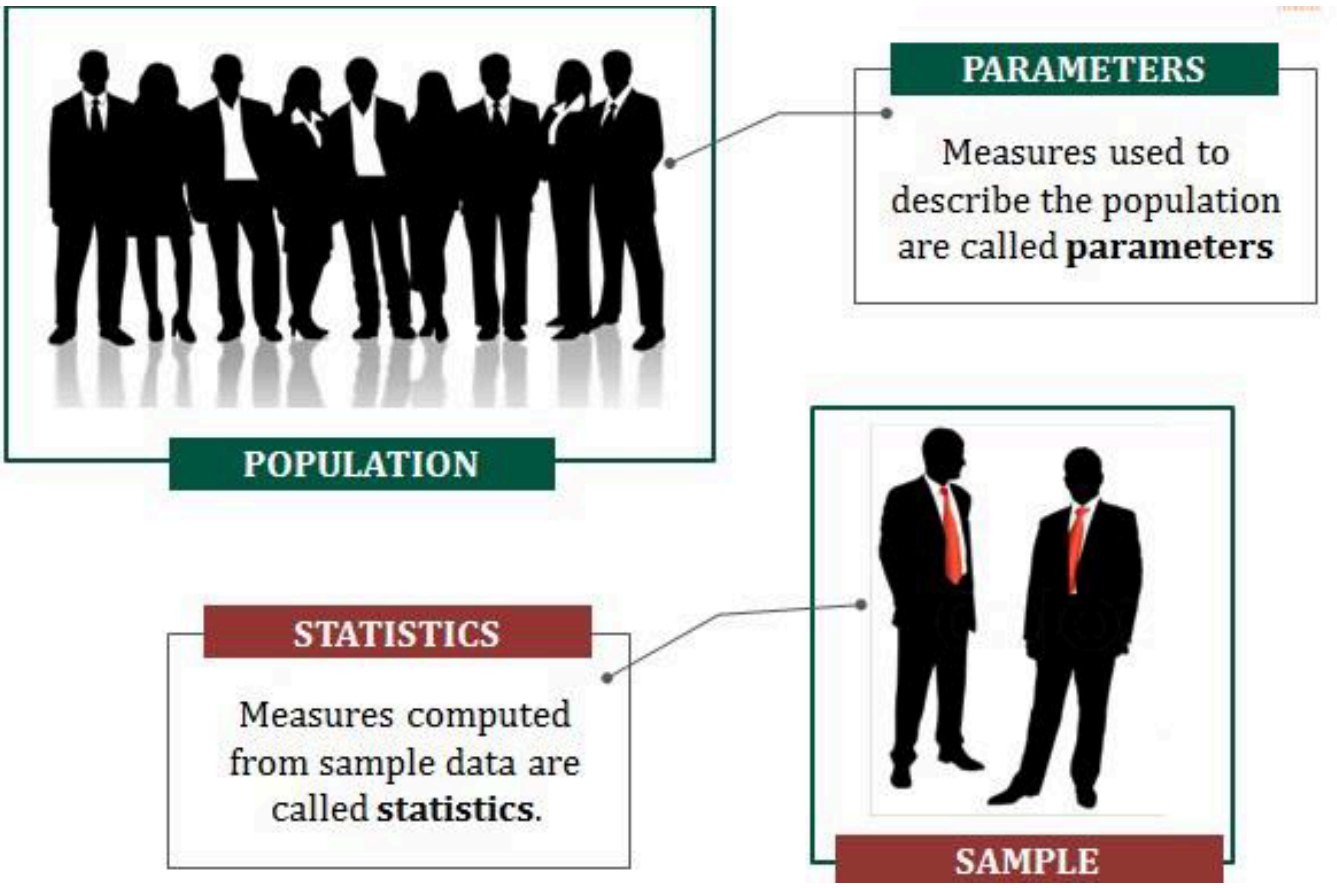
Σπανίως τα στάδια είναι πάνω από τρία.

---

5. Πληθυσμιακές παράμετροι και δειγματικά μεγέθη

Τι κάνουμε στην **επαγωγική στατιστική** (inferential statistics):

- Αναλύουμε ένα δείγμα (sample), που έχει επιλεγεί από τον πληθυσμό (population) με δειγματοληψία που να το κάνει όσο πιο αντιπροσωπευτικό (representative) γίνεται
- Υπολογίζουμε τις τιμές κάποιου **δειγματικού μεγέθους** (sample statistic)
    - Π.χ. μέση τιμή του δείγματος
- Με βάση την τιμή του δειγματικού μεγέθους, προσπαθούμε να προβλέψουμε, σε συγκεκριμένο **επίπεδο εμπιστοσύνης** (confidence level), την τιμή της αντίστοιχης **πληθυσμιακής παραμέτρου** (population parameters)
    - Π.χ. μέση τιμή πληθυσμού



Οι πληθυσμιακές παράμετροι έχουν σταθερή τιμή, που όμως είναι άγνωστη στον ερευνητή (γιατί δεν γίνεται να καταγράψουμε όλο τον πληθυσμό)

Αντίθετα, η τιμή των δειγματικών μεγεθών κυμαίνεται από δείγμα σε δείγμα

Parameter VS Statistic

When studying statistics, you might come across many terms that, if you aren't fully concentrated, will create a lot of confusion. For example, take parameter and statistic.

DEFINITION

If you are talking about a PARAMETER, you are talking about the whole population.

EXAMPLES

- In 2010, the population of the town numbered about 5000.
- This city has a population of more than 1000000.
- There are 305 doctors in this hospital.
- The number of students enrolled at the college is 15000.

DEFINITION

A STATISTIC describes only a sample of the population.

EXAMPLES

- According to a survey of 606 city residents, garbage collection was the city service people liked most.
- A survey of 2000 federation members had shown that 48% believed police should have the right to take industrial action.

Because a parameter is found out only when you know data about everyone in the population, it's fixed. In contrast, a statistic that describes the same population can vary. This is because you can take different samples from the same population and thus get different results. So, a parameter is obviously more reliable than a statistic. Still, when a population is so large that nobody has the time or the resources to ask everyone, a statistic will provide enough information to draw conclusions.

Μια ακόμα σημαντική έννοια στη στατιστική είναι η **αξιοπιστία** (reliability)

- Η αξιοπιστία χαρακτηρίζει τη συνέπεια και την ευστάθεια ενός στατιστικού μεγέθους

    - Εάν η τιμή ενός δειγματικού μεγέθους δεν αλλάζει πολύ από δείγμα σε δείγμα, τότε το δειγματικό αυτό μέγεθος χαρακτηρίζεται ως αξιόπιστο
    - Από δείγμα σε δείγμα θα υπάρχει βέβαια κάποια διακύμανση των τιμών οποιουδήποτε στατιστικού μεγέθους
- Ένα καθημερινό παράδειγμα:
    - Έχετε μια ζυγαριά μπάνιου για να μετρήσετε το βάρος σας
    - Ανεβαίνετε στη ζυγαριά τρεις φορές, και κάθε φορά σας δίνει μια ελαφρώς διαφορετική μέτρηση: 70 κιλά, 71 κιλά, και 69 κιλά
    - Αν η ζυγαριά ήταν αξιόπιστη, θα σας έδινε παρόμοιες μετρήσεις κάθε φορά που την χρησιμοποιείτε υπό τις ίδιες συνθήκες
    - Σε αυτή την περίπτωση, οι μετρήσεις δεν είναι πολύ συνεπείς, οπότε η ζυγαριά δεν μπορεί να χαρακτηριστεί αξιόπιστη

Η αξιοπιστία και η αντιπροσωπευτικότητα είναι και οι δύο σημαντικές έννοιες για την αξιολόγηση της εγκυρότητας μιας στατιστικής ανάλυσης

- Ένα στατιστικό μέγεθος μπορεί να φαίνεται αξιόπιστο σε διαφορετικά δείγματα
- Εάν όμως αυτά τα δείγματα δεν είναι αντιπροσωπευτικά του πληθυσμού, το στατιστικό μέγεθος ενδέχεται τελικά να μην είναι αξιόπιστο για την εξαγωγή ακριβών συμπερασμάτων σχετικά με την παράμετρο του πληθυσμού

Μια πληθυσμιακή παράμετρος είναι πιο αξιόπιστη από ένα δειγματικό μέγεθος

- Παρόλα αυτά, σχεδόν ποτέ δεν έχουμε το χρόνο ή τους πόρους που απαιτούνται για να καταγράψουμε έναν ολόκληρο πληθυσμό

Αντίστοιχα, ένα δειγματικό μέγεθος (π.χ. η μέση τιμή) ενός μεγαλύτερου δείγματος θα είναι πιο αξιόπιστο από το αντίστοιχο δειγματικό μέγεθος ενός μικρότερου δείγματος

- Γι αυτό επιλέγουμε σχεδόν πάντα το μεγαλύτερο δυνατό δείγμα που μπορούμε να αντέξουμε από πλευράς χρόνου, κόστους, και άλλων απαιτούμενων πόρων

Περισσότερα για την αξιοπιστία αλλά και την εμπιστοσύνη θα πούμε αργότερα

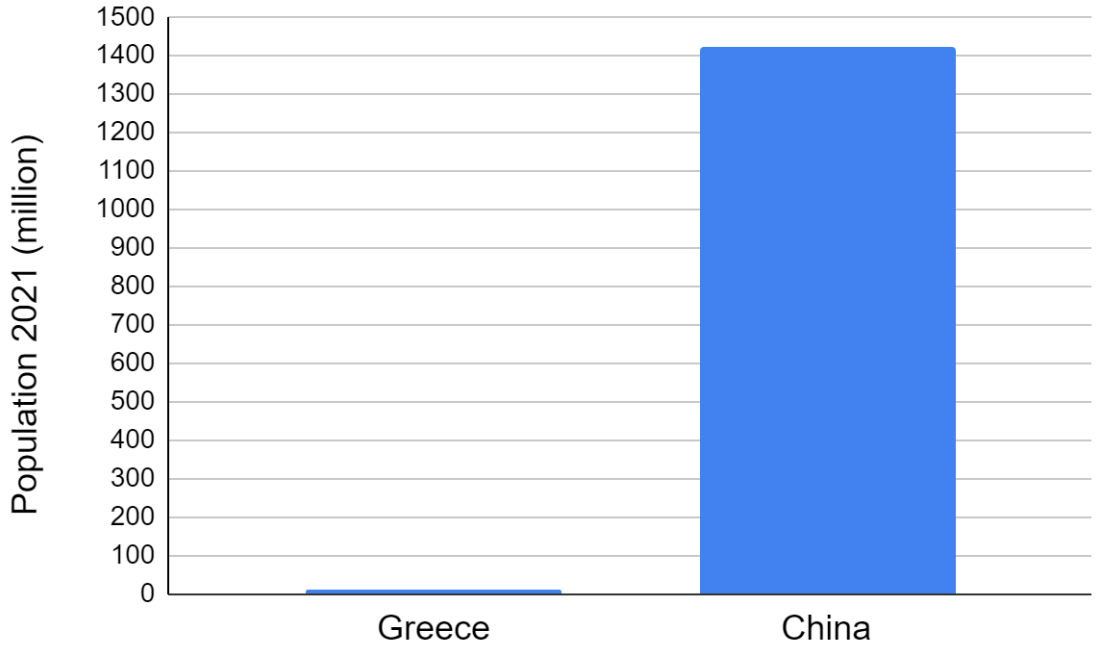
5.1. Ο νόμος των μεγάλων αριθμών

Ο **νόμος των μεγάλων αριθμών** (law of large numbers) ορίζει ότι όσο μεγαλώνει το (απόλυτο) μέγεθος ενός δείγματος (π.χ. 10 άτομα, 50 άτομα, 200 άτομα), τόσο τα δειγματικά μεγέθη (π.χ. ο μέσος όρος του δείγματος) θα προσεγγίζουν την τιμή των αντίστοιχων πληθυσμιακών παραμέτρων (π.χ. τον μέσο όρο του πληθυσμού)

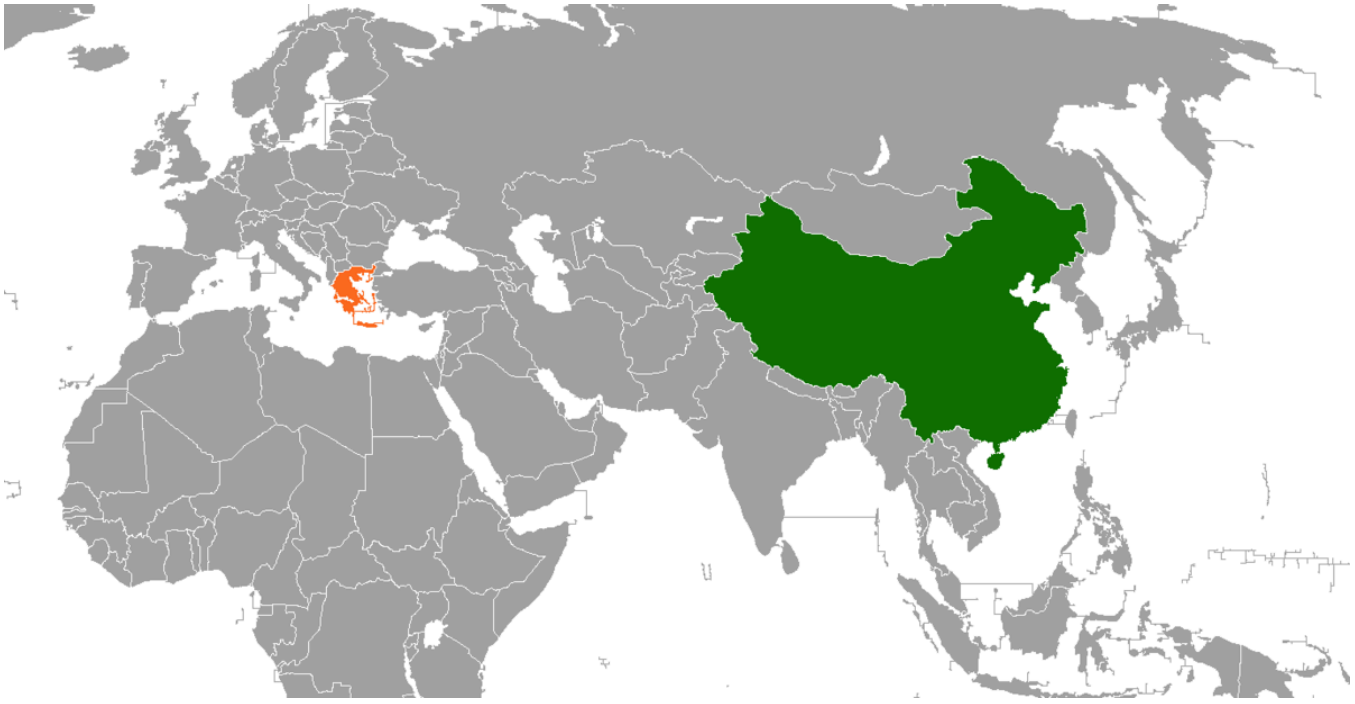
- Και αυτό εξαρτάται μόνο από το απόλυτο μέγεθος του δείγματος (π.χ. 20 άτομα) και όχι από το μέγεθος του δείγματος ως ποσοστό του μεγέθους του πληθυσμού (π.χ. 10% του πληθυσμού, εάν ο πληθυσμός είχε 200 άτομα)!

Ας δούμε ένα παράδοξο παράδειγμα, που προκύπτει από το νόμο αυτό:

- Πληθυσμός Ελλάδας (2021) = 10.64 εκατομμύρια
- Πληθυσμός Κίνας (2021) = 1.412 δισεκατομμύρια = 1421 εκατομμύρια



Εάν επιλεγεί ένα δείγμα, ας πούμε 10 χιλιάδων ατόμων, από την Ελλάδα, και ένα άλλο δείγμα επίσης 10 χιλιάδων ατόμων από την Κίνα, ποιο θα είναι πιο αξιόπιστο;



Και τα δύο θα είναι το ίδιο αξιόπιστα!

- Η αξιοπιστία, λόγω του νόμου των μεγάλων αριθμών, εξαρτάται από το απόλυτο μέγεθος του δείγματος (10 χιλιάδες), όχι από το μέγεθος του πληθυσμού από τον οποία ελήφθη (10.64 εκατομμύρια για την Ελλάδα, 1421 εκατομμύρια για την Κίνα)!
- Το ότι το δείγμα των 10 χιλιάδων ατόμων αποτελεί το 10,000 / 10,640,000 × 100 = **0.094%** της Ελλάδας, και το 10,000 / 1,412,000,000 × 100 = **0.00071%** της Κίνας, δεν έχει καμία σημασία
- Το μόνο που έχει σημασία είναι ότι και τα δύο δείγματα έχουν το ίδιο απόλυτο μέγεθος, δηλαδή 10,000 άτομα!

5.2. Δειγματοληπτικό και στατιστικό σφάλμα

Μη δειγματοληπτικό σφάλμα (nonsampling error), για παράδειγμα

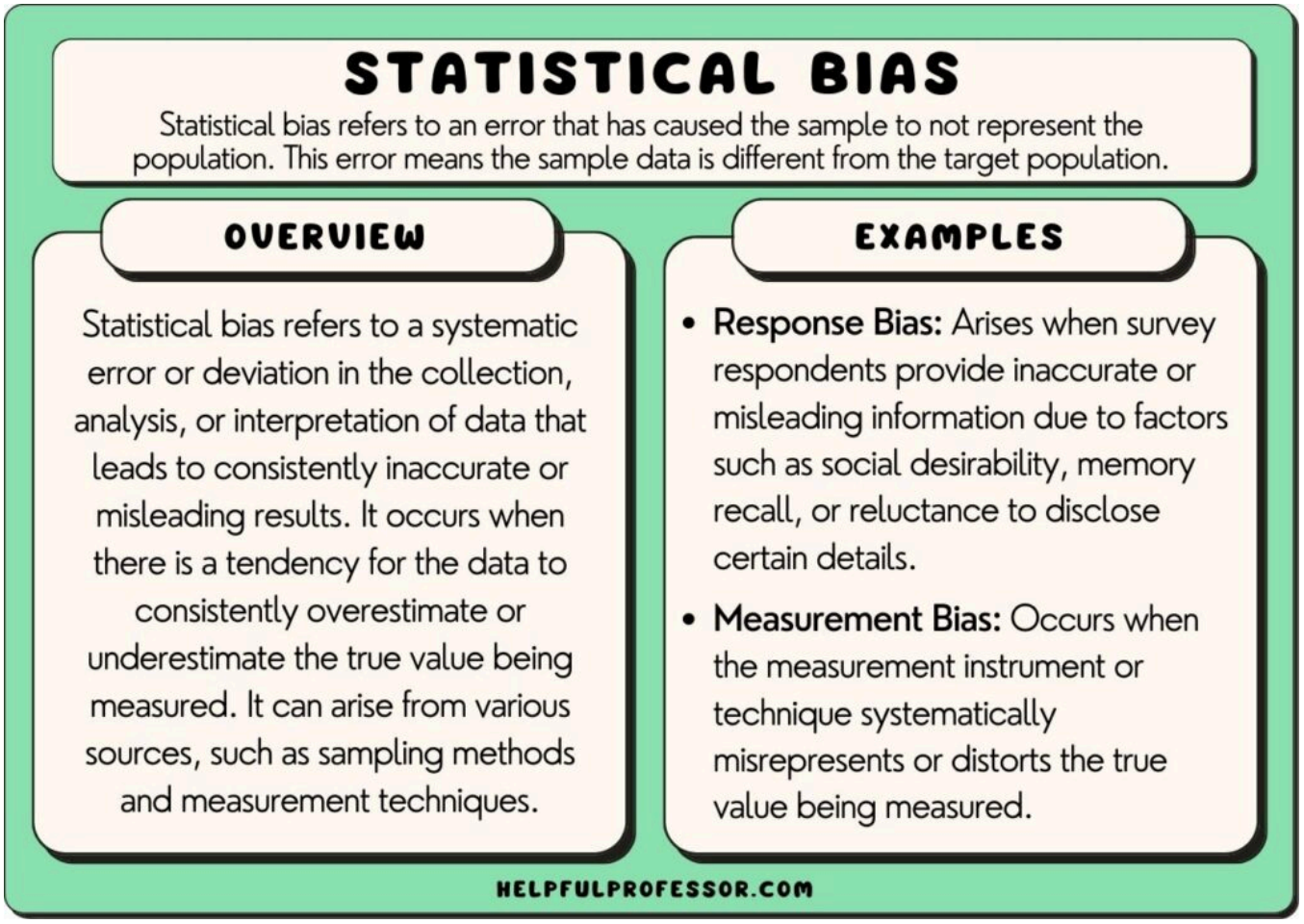
- Επιλέγεται ακατάλληλο δείγμα
- Γίνονται λάθη κατά τη μέτρηση ή συλλογή των δεδομένων

Δειγματοληπτικό σφάλμα (sampling error)

- Όταν το στατιστικό μέγεθος (sample statistic) διαφέρει από την πληθυσμιακή παράμετρο (population parameter)
- Αυτό μπορεί να συμβεί στη δειγματοληψία ακόμα και όταν όλα έχουν σχεδιαστεί σωστά
    - Οφείλεται στην τυχαία διακύμανση που παρατηρείται από το ένα δείγμα σε άλλο (luck of the draw)

Στατιστικό σφάλμα (bias)

- Στατιστικό σφάλμα έχουμε όταν το δείγμα δεν είναι αντιπροσωπευτικό του πληθυσμού
- Το στατιστικό σφάλμα όμως σχετίζεται με την έννοια του **συστηματικού σφάλματος** (systematic bias), το οποίο διαφέρει από το **τυχαίο σφάλμα** (random error)
- Η έννοια του στατιστικού σφάλματος είναι εξαιρετικά σημαντική στη στατιστική



Από την wikipedia:

*"An example of **systematic bias** would be the bias of a thermometer that always reads three degrees colder than the actual temperature because of an incorrect initial calibration or labeling, whereas one that gave random values within five degrees either side of the actual temperature would be considered a **random error**."*

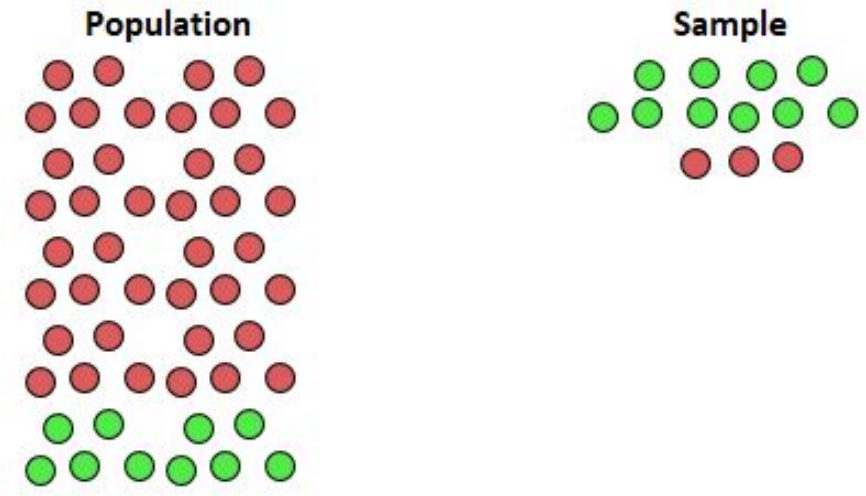
Περισσότερες λεπτομέρειες από την wikipedia:

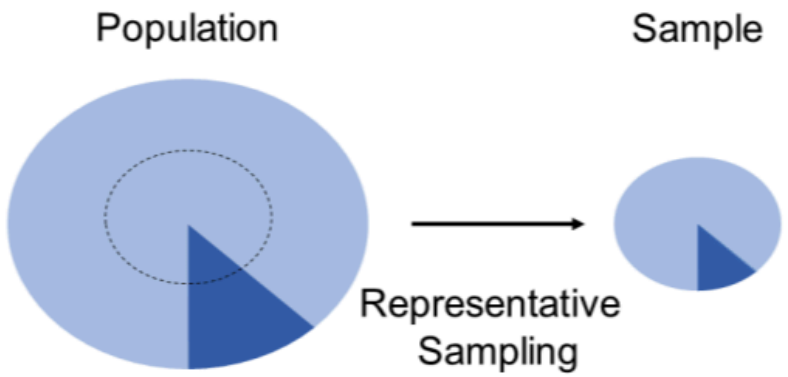
*In statistics, the term **bias** is used for describing several different concepts:*

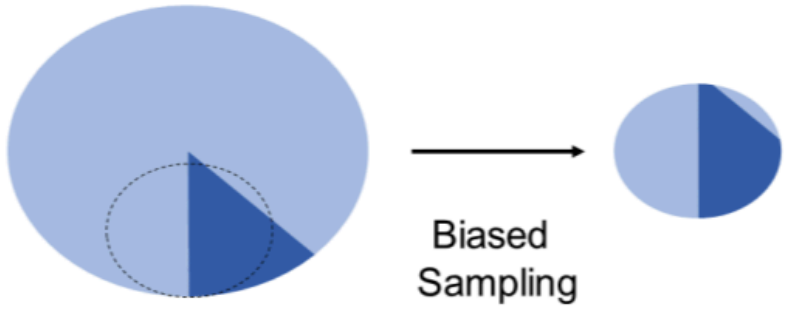
- *A **biased sample** is one in which some members of the population are more likely to be included than others.*
- *The **bias of an estimator** is the difference between an estimator's expectation and the true value of the parameter being estimated.*
- ***Omitted-variable bias** is the bias that appears in estimates of parameters in a regression analysis when the assumed specification is incorrect, in that it omits an independent variable that should be in the model.*
- *In statistical hypothesis testing, a test is said to be unbiased when the probability of rejecting the null hypothesis exceeds the significance level when the alternative is true and is less than or equal to the significance level when the null hypothesis is true.* (Αυτό δεν θα γίνει κατανοητό τώρα)
- ***Systematic bias** or **systemic bias** are external influences that may affect the accuracy of statistical measurements.*

Κάποια από τα παραπάνω θα συζητηθούν αργότερα στο μάθημα.

Παραδείγματα εσφαλμένων (μη αντιπροσωπευτικών δειγμάτων)







Άλλα παραδείγματα μη αντιπροσωπευτικών δειγμάτων:

- Αν ρωτούσα τους φοιτητές που κάθονται στο πρώτο θρανίο κατά πόσον τους αρέσει το μάθημα, με σκοπό να εξάγω συμπεράσματα για τους φοιτητές όλου του έτους
    - Ίσως οι φοιτητές που επιλέγουν να κάτσουν στο πρώτο θρανίο είναι πιο μελετηροί
    - Σκεφτείτε επίσης ότι δεν έρχονται όλοι οι φοιτητές στις διαλέξεις

- Αν ρωτούσα την παρούσα τάξη (πρωτοετείς, ΔΕΣ) τι τους ικανοποιεί και τι όχι από το ΠΑΠΕΙ, με σκοπό να εξάγω συμπεράσματα για όλους τους φοιτητές όλων των ετών και όλων των τμημάτων του ΠΑΠΕΙ
    - Οι πρωτοετείς φοιτητές του ΔΕΣ δεν θα είναι αντιπροσωπευτικοί όλων των άλλων τμημάτων και ετών του πανεπιστημίου
    - Θα έπρεπε να πάρω ένα δείγμα που να περιέχει φοιτητές από όλα τα τμήμα και όλα τα έτη

---

6. Δεδομένα (data)

Είδη δεδομένων

1. **Ποσοτικά** (quantitative)
    - Αριθμοί (numbers), πλήθος (counts), μετρήσεις (measurements)
        - Π.χ. ύψος, βάρος, μηνιαίος μισθός
2. **Ποιοτικά** (qualitative)
    - Κατηγορίες (categories), χαρακτηριστικά (attributes)
        - Φύλο (άνδρας/γυναίκα), ομάδα (Ολυμπιακός/Παναθηναϊκός/ΑΕΚ κλπ.), κόμμα που ψηφίζετε

Εξαιρετικά σημαντικό να ξέρουμε τα εξής για τα δεδομένα μας

1. **Πηγή** (source) προέλευσης
2. **Χρονολογία** (π.χ. έτος) στην οποία αναφέρονται
3. **Μονάδες** μέτρησης (units)

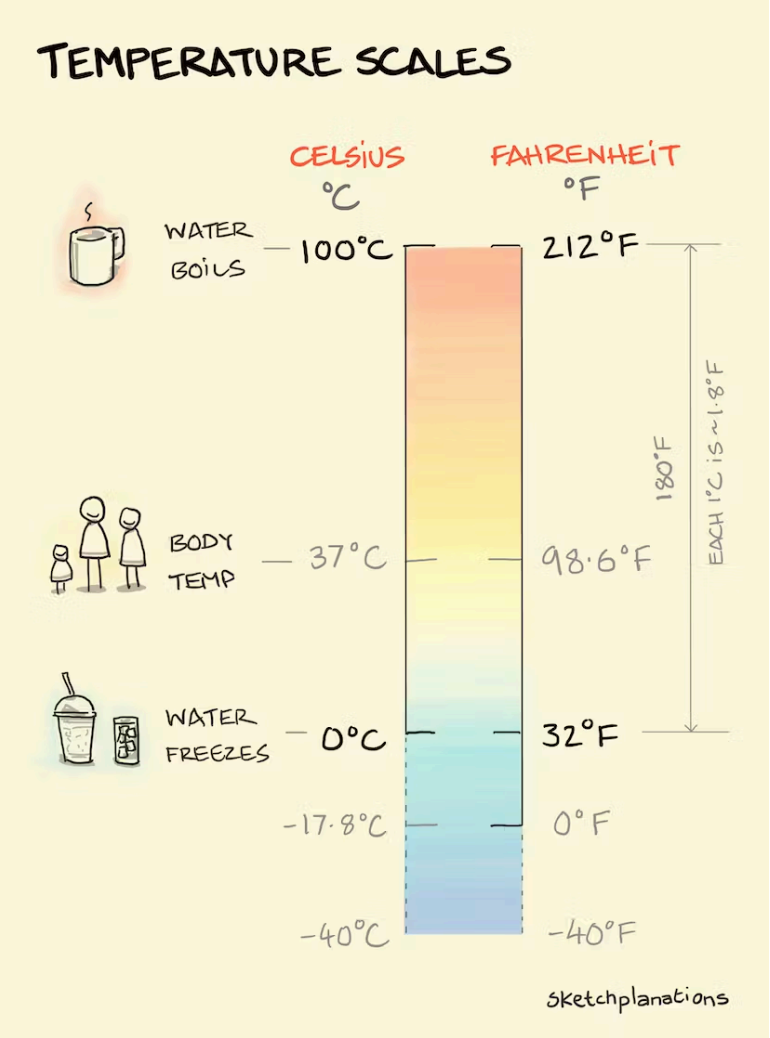
6.1. Ποσοτικά δεδομένα

Είδη ποσοτικών δεδομένων

1. **Διακριτά** (discrete)
    - Ακέραιοι αριθμοί
        - Π.χ. αριθμός παιδιών μιας οικογένειας, αριθμός φοιτητών σε ένα έτος
2. **Συνεχή** (continuous)
    - Πραγματικοί αριθμοί
        - Π.χ. ημερήσιες εκπομπές διοξειδίου του άνθρακα (CO_2) από ένα εργοστάσιο, ετήσια παραγωγή γάλακτος από την κτηνοτροφία μιας περιοχής, ετήσιο εισόδημα ενός επαγγελματία

Μια ακόμα διάκριση ποσοτικών δεδομένων

1. **Διαστήματος** (interval)
    - Τα δεδομένα διαστήματος είναι αριθμητικά αλλά, από τη φύση τους, δεν έχουν σημείο μηδενισμού (δηλαδή δεν υπάρχει σημείο στο οποίο να μην υπάρχει καθόλου η μετρούμενη ποσότητα)
        - Π.χ. το έτος (π.χ. 10000 π.Χ, 1972 μ.Χ., 2024 μ.Χ.) ή η θερμοκρασία (°C ή °F)



2. **Αναλογίας** (ratio)
    - Σημαντική διαφορά τους από τα δεδομένα διαστήματος είναι ότι περιέχουν το σημείο του μηδενός
    - Ως εκ τούτου, έχει έννοια να εξετάζονται αναλογίες (ratios) αυτών των δεδομένων
        - Για παράδειγμα, ένας ενήλικας που ζυγίζει 80 κιλά είναι δύο φορές βαρύτερος από ένα παιδί που ζυγίζει 40 κιλά
            - Ενώ 40°C (=104°F) δεν είναι δύο φορές 20°C (=68°F)
        - Ένα άλλο παράδειγμα είναι οι τιμές ενός προϊόντος, π.χ. €10 που είναι πράγματι διπλάσια της τιμής των €5

- Παράλληλα δε, μηδέν ευρώ σημαίνει ότι το προϊόν διατίθεται δωρεάν, δηλαδή η τιμή του μηδενός έχει φυσικό νόημα (σε αντίθεση, π.χ. με τη θερμοκρασία)

6.2. Ποιοτικά δεδομένα

Είδη ποιοτικών δεδομένων

1. **Ονομαστικά** (nominal) είναι τα ποιοτικά δεδομένα που δεν επιδέχονται ταξινόμηση
    - Ονόματα, ετικέτες/ταμπέλες (labels), κατηγορίες
        - Π.χ. , το χρώμα που σας αρέσει, η ομάδα ποδοσφαίρου που στηρίζετε, οι απαντήσεις ναι/όχι/δεν γνωρίζω ενός ερωτηματολογίου
2. **Τακτικά** (ordinal) είναι τα ποιοτικά δεδομένα που μπορούν να ταξινομηθούν
    - Ένα παράδειγμα είναι οι βαθμίδες καθηγητών στα πανεπιστήμια, από την χαμηλότερη στην υψηλότερη: Λέκτορας (lecturer, δεν υπάρχει πλέον αυτή η βαθμίδα), Επίκουρος Καθηγητής (Assistant Professor), Αναπληρωτής Καθηγητής (Associate Professor), Καθηγητής (Professor)
    - Άλλα παραδείγματα: οι βαθμοί που δίνονται σε Αμερικανικά πανεπιστήμιο (A: άριστα, B: πολύ καλά, C: καλά, D: βάση, F: αποτυχία), οι βαθμοί των αξιωματικών στο στρατό (Στρατιώτης, Υποδεκανέας, Δεκανέας, Λοχίας, Επιλογίας, Αρχιλογίας, κλπ.)

Προσοχή:

- Μερικές φορές μπορεί για ευκολία να καταχωρούμε ποιοτικά δεδομένα με αριθμητικές τιμές, π.χ. άνδρας=1, γυναίκα=2
    - Αυτό δεν σημαίνει όμως ότι η γυναίκα που καταχωρείται με την τιμή 2 είναι δύο φορές “μεγαλύτερη” από τον ανδρα

Ανακεφαλαίωση των διαφορετικών ειδών δεδομένων με περισσότερα παραδείγματα

Table 1-1 Levels of Measurement of Data

Level	Summary	Example	
Nominal	Categories only. Data cannot be arranged in an ordering scheme.	Student states: 5 Californians 20 Texans 40 New Yorkers	Categories or names only.
Ordinal	Categories are ordered, but differences can't be found or are meaningless.	Student cars: 5 compact 20 mid-size 40 full size	An order is determined by "compact, mid-size, full-size."
Interval	Differences are meaningful, but there is no natural starting point and ratios are meaningless.	Campus temperatures: 5°F 20°F 40°F	0°F doesn't mean "no heat." 40°F is not twice as hot as 20°F.
Ratio	There is a natural zero starting point and ratios are meaningful.	Student commuting distances: 5 mi 20 mi 40 mi	40 mi is *twice* as far as 20 miles.

---

7. Μέτρα θέσεως και διασποράς

7.1. Πλήθος, ελάχιστο, μέγιστο, και εύρος

Θεωρούμε το ακόλουθο σύνολο δεδομένων

1, 3, 3, 5, 6, 6, 6, 7, και 9

Αν και δεν μας απασχολεί κατά πόσον αυτά τα δεδομένα προέρχονται από κάποιον πληθυσμό, θα αναφέρομαι σε αυτά σαν δείγμα (sample)

Αν και πρέπει πάντα να γνωρίζουμε τη φύση και τις μονάδες μέτρησης των δεδομένων, στην προκειμένη περίπτωση προχωράμε χωρίς να το εξετάζουμε αυτό

- Δεν μας ενδιαφέρει (για εκπαιδευτικούς σκοπούς) αν τα δεδομένα αντιπροσωπεύουν κάτι

- Θεωρούμε ότι δεν έχουν μονάδες μέτρησης, δηλαδή ότι είναι αδιάστατα (dimensionless)

Πλήθος δεδομένων (δηλαδή το **μέγεθος** του δείγματος, sample size, που συνήθως παριστάνεται με N) = 9

- Το δείγμα έχει 9 **παρατηρήσεις** (cases ή observations)

Τα δεδομένα είναι ήδη ταξινομημένα από το μικρότερο (ελάχιστο ή minimum) ως το μεγαλύτερο (μέγιστο ή maximum), επομένως εύκολα μπορούμε να δούμε ότι

- Ελάχιστη τιμή ή **ελάχιστο** (minimum) = 1
- Μέγιστη τιμή ή **μέγιστο** (maximum) = 9
- **Εύρος** (range) = Μέγιστο − Ελάχιστο = 9 − 1 = 8

7.2 Πίνακες συχνοτήτων και ιστογράμματα

Φτιάχνουμε ένα πίνακα συχνοτήτων προκειμένου μετά να σχεδιάσουμε ένα ιστόγραμμα, παριστάνοντας με Case τον αύξοντα αριθμό των παρατηρήσεων (case number) και με X τις τιμές των δεδομένων

Αρχικά τα δεδομένα μπορούν να παρασταθούν σε μορφή πίνακα ως εξής:

Case	*X*
1	1
2	3
3	3
4	5
5	6
6	6
7	6
8	7
9	9

Ακολούθως τα ομαδοποιούμε στον ακόλουθο **πίνακα συχνοτήτων** (frequency table), που για τα συγκεκριμένα δεδομένα έχει μόνο ακέραιες τιμές:

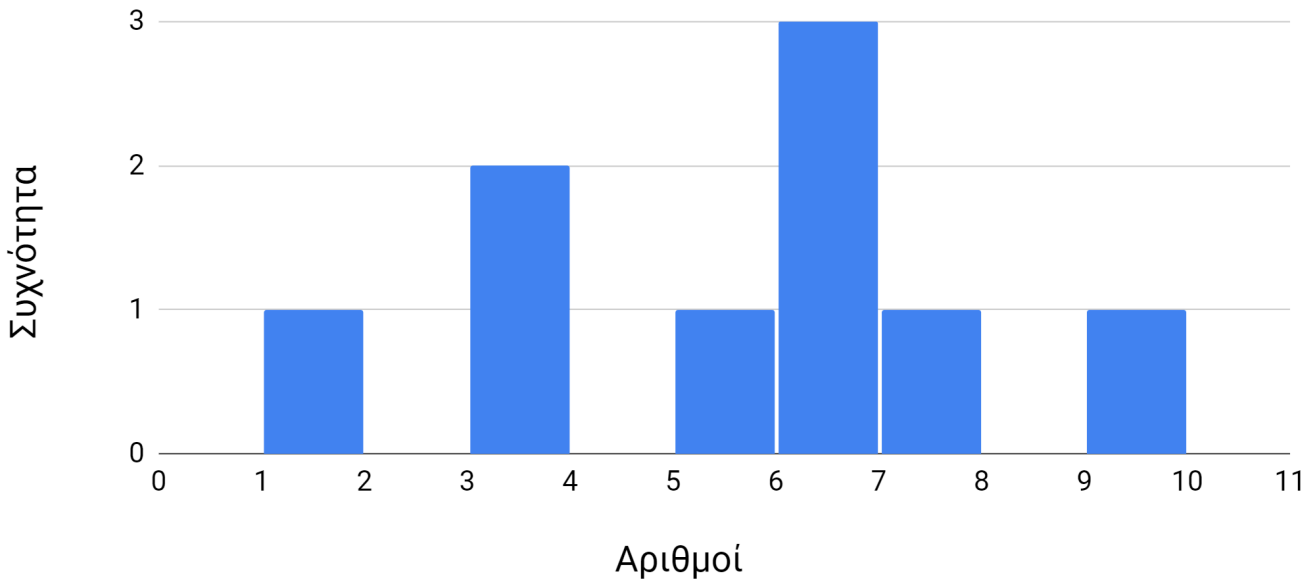
Τιμή (value)	*Συχνότητα (frequency)*	*%*
1	1	11.1
3	2	22.2
5	1	11.1
6	3	33.3
7	1	11.1

9	1	11.1
Σύνολο	9	100

Προφανώς

- Οι συχνότητες αθροίζουν σε 9, δηλαδή όσα τα δεδομένα του δείγματος
- Τα αντίστοιχα ποσοστά αθροίζουν σε 100%
    - *(Αν είστε παρατηρητικοί θα προσέξετε ότι η δεξιά στήλη δεν αθροίζει σε 100% αλλά σε 99.9% λόγων στρογγυλοποίησης)*

Το αντίστοιχο ιστόγραμμα (Google Sheets) έχει ως ακολούθως:



(Κανονικά πρέπει να σχεδιάζεται και η γραμμή του κάθετου άξονα, αυτό όμως δεν είναι δυνατόν στο Google Sheets)

Για να καταλάβουμε τις τιμές του οριζόντιου άξονα, πρέπει να φανταστούμε ότι ο πίνακας συχνοτήτων στην πραγματικότητα αποτελείται από τάξεις ή κλάσεις (classes), η κάθε μια εκ των οποίων έχει κάτω και άνω όριο

- Οι κλάσεις κατά κανόνα έχουν το ίδιο πλάτος
- Επίσης κατά κανόνα υποθέτουμε ότι ως προς το κάτω όριο οι κλάσεις είναι κλειστές (και αυτό συμβολίζεται με αριστερή αγκύλη), ως προς το πάνω δε ανοιχτές (και αυτό συμβολίζεται με δεξιά παρένθεση)
    - Κλειστό άκρο κλάσης σημαίνει ότι η κλάση περιλαμβάνει την τιμή του άκρου
    - Ανοικτό άκρο κλάσης σημαίνει ότι η κλάση δεν περιλαμβάνει την τιμή του άκρου

Κλάση (class)	*Συχνότητα (frequency)*	%
[1–2)	1	11.1
[2–3)	0	0
[3–4)	2	22.2

[4–5)	0	0
[5–6)	1	11.1
[6–7)	3	33.3
[7–8)	1	11.1
[8–9)	0	0
[9–10)	1	11.1
Σύνολο	9	100

Στον παραπάνω πίνακα, γράψαμε το “[” για να δείξουμε το κλειστό αριστερό άκρο και “)” για να δείξουμε το ανοιχτό δεξιό άκρο της κάθε κλάσης

- Κανονικά όμως αυτά δεν φαίνονται σε πίνακες συχνοτήτων
- Μερικές φορές γράφουμε τα όρια των κλάσεων έτσι ώστε να είναι ξεκάθαρο σε ποια κλάση περιλαμβάνεται το κάθε δεδομένο

Κλάση (class)	*Συχνότητα (frequency)*	%
1–1.999	1	11.1
2–2.999	0	0
3–3.999	2	22.2
4–4.999	0	0
5–5.999	1	11.1
6–6.999	3	33.3
7–7.999	1	11.1
8–8.999	0	0
9–9.999	1	11.1
Σύνολο	9	100

7.3. Τεταρτημόρια, διάμεσος, και επικρατούσα τιμή

Στον παρακάτω πίνακα απεικονίζονται τα μήκη των υποβρυχίων (U-boats) τύπου IX της Ναζιστικής Γερμανίας, κατά τον Δεύτερο Παγκόσμιο Πόλεμο.

Παρατήρηση	Τύπος	Μήκος (m)
1	IX-A	75.5
2	IX-B	76.5
3	IX-C	76.8
4	IX-C/40	76.8
5	IX-D2	87.6
6	XXI	76.7

Το μέγεθος του δείγματος είναι N = 6

Θα υπολογίσουμε τα ακόλουθα:

- Ελάχιστο (min)
- **Διάμεση** τιμή (median)
- Μέγιστο (max)
- **Επικρατούσα τιμή** (mode)
    - Η πιο συχνή τιμή (δηλαδή η τιμή εκείνη που εμφανίζεται περισσότερες φορές)
    - Η επικρατούσα τιμή μπορεί και να μην υπάρχει (αν κανένα από τα δεδομένα δεν εμφανίζεται πάνω από μια φορά ή περισσότερες φορές από τα άλλα)
    - Η τιμή που είναι στη μέση των δεδομένων, δηλαδή κάτω από την οποία βρίσκεται το 50% των δεδομένων
    - Η διάμεση τιμή είναι το δεύτερο τεταρτημόριο
- **Πρώτο τεταρτημόριο** (first quartile ή Q_1)
    - Η τιμή κάτω από την οποία βρίσκεται το 25% των δεδομένων
    - Το πρώτο τεταρτημόριο είναι η διάμεσος του πρώτου μισού!
- **Τρίτο τεταρτημόριο** (third quartile ή Q_3)
    - Η τιμή κάτω από την οποία βρίσκεται το 75% των δεδομένων
    - Το τρίτο τεταρτημόριο είναι η διάμεσος του δεύτερου μισού!
- **Ενδοτεταρτημοριακό εύρος** (interquartile range ή IQR)
    - Ισούται με $Q_3 - Q_1$

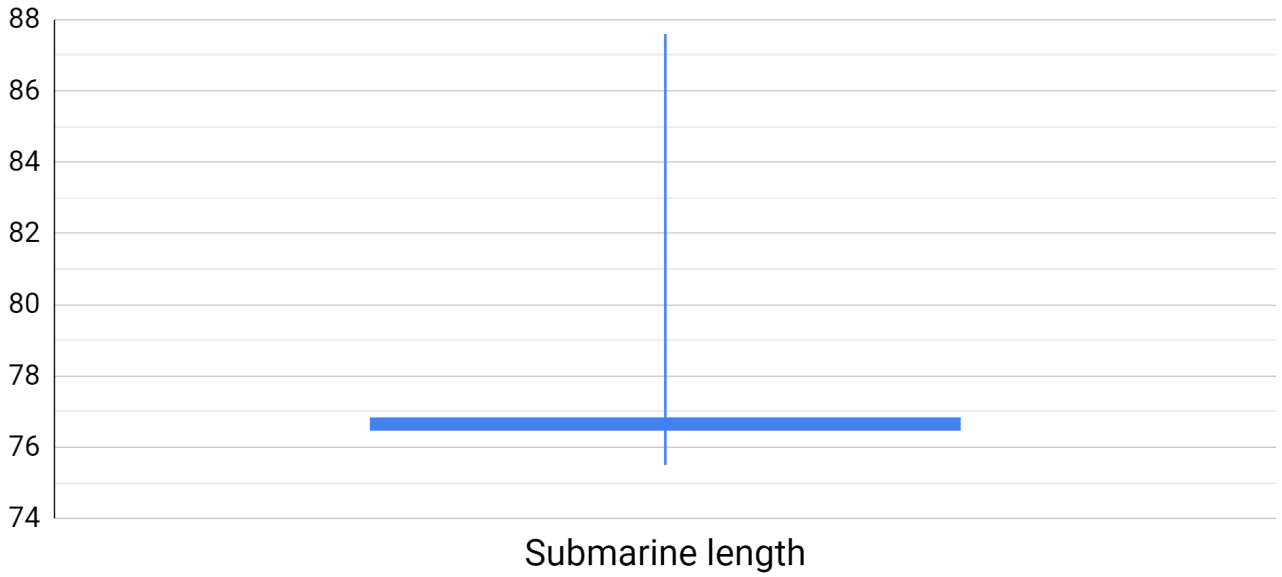
Για να υπολογίσουμε τα τεταρτημόρια και τη διάμεσο, πρέπει να ταξινομήσουμε τα δεδομένα, από το μικρότερο στο μεγαλύτερο

Πρωτογενή (αταξινόμητα) δεδομένα	Παρατήρηση	Ταξινομημένα δεδομένα
75.5	1	75.5
76.5	2	76.5
76.8	3	76.7
76.8	4	76.8
87.6	5	76.8
76.7	6	87.6

Δουλεύοντας στα ταξινομημένα δεδομένα, υπολογίζουμε τα παρακάτω:

- Ελάχιστο = 75.5
- Μέγιστο = 87.6
- Εύρος = Μέγιστο − Ελάχιστο = 87.6 − 75.5 = 12.1
- Διάμεσος = (76.6+76.8)/2 = 76.7
    - Επειδή το μέγεθος του δείγματος είναι ζυγός αριθμός (6), καμία παρατήρηση δεν είναι ακριβώς στη μέση
    - Άρα παίρνουμε το ημιάθροισμα των δύο μεσαίων παρατηρήσεων
- Επικρατούσα τιμή = 76.8
    - Το 76.8 είναι η μόνη τιμή που εμφανίζεται δύο φορές
- Πρώτο τεταρτημόριο Q_1 = 76.5
    - Το πρώτο μισό περιέχει τις παρατηρήσεις 75.5, 76.5 και 76.7
    - Η διάμεσος αυτών των τριών αριθμών, είναι το 76.5
- Τρίτο τεταρτημόριο Q_3 = 76.8
    - Το δεύτερο μισό περιέχει τις παρατηρήσεις 76.8, 76.8 και 87.6
    - Η διάμεσος αυτών των τριών αριθμών, είναι το 76.8
- Ενδοτεταρτημοριακό εύρος IQR = $Q_3 - Q_1$ = 76.8 − 76.5 = 0.3

Το ελάχιστο, το Q1, η διάμεσος, το Q3, και το μέγιστο αναφέρονται και ως five number summary, και απεικονίζονται σε διάγραμμα boxplot

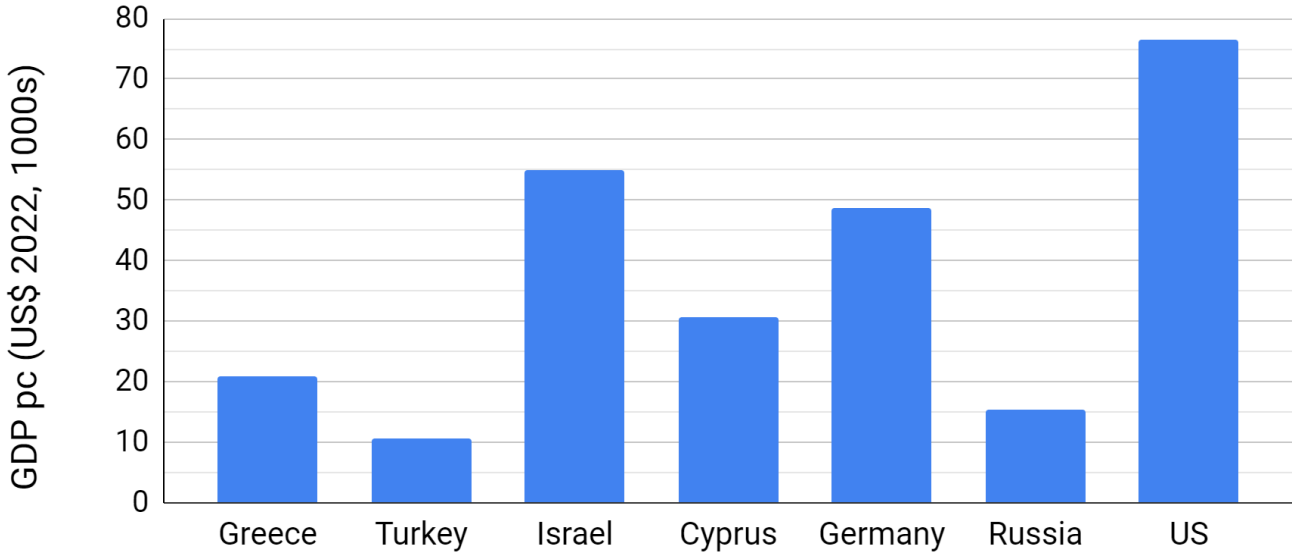


Η εμφάνιση του παραπάνω boxplot είναι λίγο παραπλανητική γιατί το ενδοτεταρτημοριακό πλάτος είναι πολύ μικρό

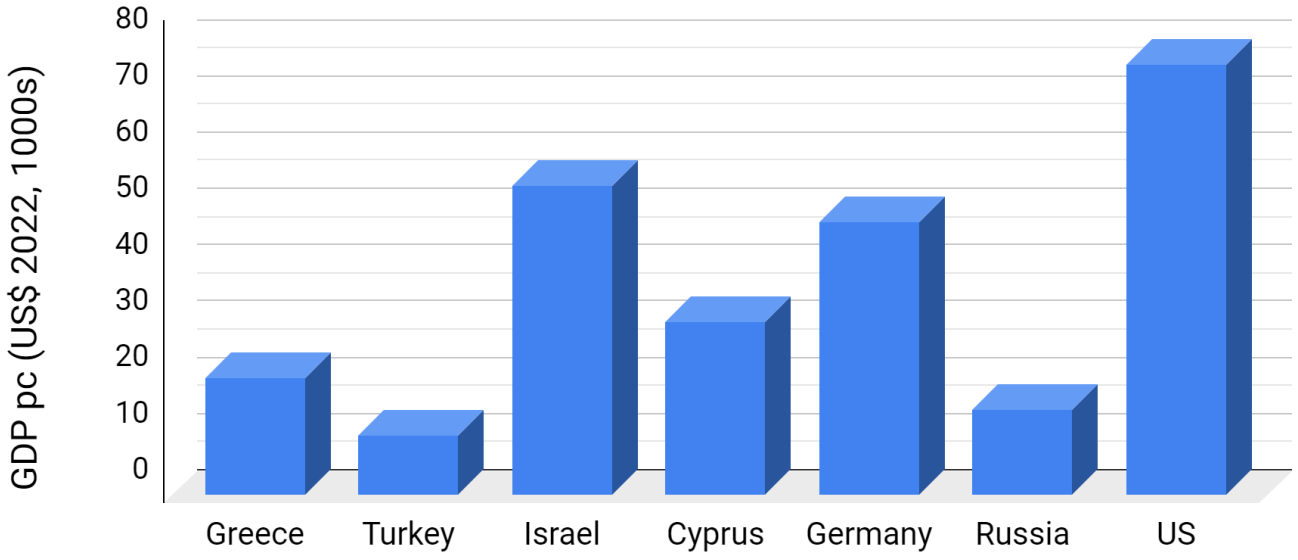
Ας αναλύσουμε και ένα άλλο σύνολο 7 χωρών με το ΑΕΠ ανά κεφαλή (GDP per capita) του 2022 (σε US$)

Παρατήρηση	*Country*	*GDP pc (US$ 2022, 1000s)*
1	Greece	20.9
2	Turkey	10.6
3	Israel	55
4	Cyprus	30.7
5	Germany	48.6
6	Russia	15.3
7	US	76.4

Τα δεδομένα αυτά μπορούμε να τα απεικονίσουμε στο παρακάτω ραβδόγραμμα (bar chart)



Τρισδιάστατα διαγράμματα, όπως αυτό που ακολουθεί, πρέπει να αποφεύγονται



Ταξινομημένα δεδομένα

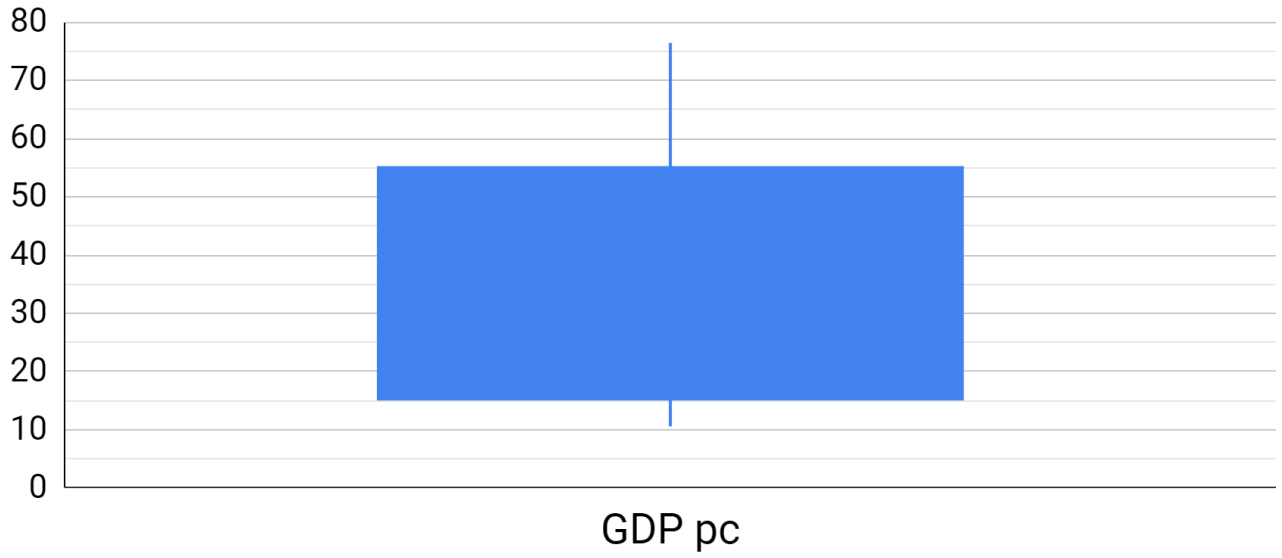
Παρατήρηση	*Country*	*GDP pc (US$ 2022, 1000s)*
1	Turkey	10.6
2	Russia	15.3
3	Greece	20.9
4	Cyprus	30.7
5	Germany	48.6
6	Israel	55
7	US	76.4

Υπολογισμοί

- Ελάχιστο = 10.6
- Μέγιστο = 76.4
- Εύρος = Μέγιστο − Ελάχιστο = 76.4 − 10.6 = 65.8
- Διάμεσος = 30.7
    - Επειδή το μέγεθος του δείγματος είναι μονός αριθμός (7), υπάρχει παρατήρηση (το 30.7) που είναι ακριβώς στη μέση
- Επικρατούσα τιμή δεν υπάρχει
- Πρώτο τεταρτημόριο Q_1 = 15.3
    - Για να μην μετρήσουμε τη διάμεσο (30.7) δύο φορές, και στο πρώτο μισό, και στο δεύτερο μισό, την αγνοούμε και θεωρούμε ότι το πρώτο μισό αποτελείται από τις πρώτες 3 παρατηρήσεις
- Τρίτο τεταρτημόριο Q_3 = 55
    - Για να μην μετρήσουμε τη διάμεσο (30.7) δύο φορές, και στο πρώτο μισό, και στο δεύτερο μισό, την αγνοούμε και θεωρούμε ότι το δεύτερο μισό αποτελείται από τις τελευταίες 3 παρατηρήσεις
- Ενδοτεταρτημοριακό εύρος IQR = $Q_3 - Q_1$ = 55 − 15.3 = 39.7

Σε επόμενο παράδειγμα θα δούμε έναν ακριβέστερο τρόπο υπολογισμού τη ς θέσης της διαμέσου, του Q_1 και του Q_3

Το boxplot αυτών των αριθμών είναι ως ακολούθως



Πως θα πρέπει να προσεγγίζουμε το boxplot

- Το κεντρικό “κουτί”(box) εκτείνεται (από κάτω προς τα πάνω) ώστε να καλύπτει το 50% των δεδομένων του δείγματος, και μάλιστα τα πιο συνηθισμένα, εκείνα που δεν έχουν πολύ μικρές ούτε πολύ μεγάλες τιμές
- Η κάτω “ουρά” (whisker, που σημαίνει μουστάκι) εκτείνεται όσο το 25% των δεδομένων που έχουν τις μικρότερες τιμές
- Η πάνω “ουρά” (whisker) εκτείνεται όσο το 25% των δεδομένων που έχουν τις μεγαλύτερες τιμές

Γι’ αυτό το boxplot ονομάζεται και box and whiskers plot

Στο παραπάνω boxplot παρατηρούμε ότι η κάτω ουρά (που αντιστοιχεί στις μικρότερες τιμές δεδομένων) είναι μικρότερη της πάνω ουράς (που αντιστοιχεί στις μικρότερες μεγαλύτερες δεδομένων)

- Αυτό σημαίνει ότι το δείγμα αυτό των 7 δεδομένων χαρακτηρίζεται από δεξιά **στρέβλωση** (skewness)
    - Θα πούμε περισσότερα για τη στρέβλωση αργότερα

7.4. Εσωτερικοί και εξωτερικοί φράχτες

Έλεγχος για ακραίες τιμές (outliers) με εσωτερικούς και εξωτερικούς φράχτες (inner και outer fences)

- Lower outer fence: $Q_1 - 3\times IQR$
- Lower inner fence: $Q_1 - 1.5\times IQR$
- *(Κάπου εδώ θα βρίσκεται η διάμεσος)*
- Upper inner fence: $Q_3 + 1.5\times IQR$
- Upper outer fence: $Q_3 + 3\times IQR$

Δεδομένα που είναι

- μέσα στους εσωτερικούς φράχτες, δεν θεωρούνται outliers

- έξω από τους εσωτερικούς φράχτες αλλά μέσα στους εξωτερικούς φράχτες, ίσως είναι outliers
- έξω από τους εξωτερικούς φράχτες, κατά πάσαν πιθανότητα είναι outliers

Παράδειγμα (από εδώ): θεωρούμε το σύνολο δεδομένων

29, 179, 180, 201, 300, 301, 304, 350, 399, 401, 455, 501, 503, 540, 543, 549, 560, 561, 562, 563, 569, 570, 599, 601, 603, 650, 701, 703, 704, 709, 713, 733, 745, 801, 900, 982, 983, 985, 999, 1001, 1002, 1003, 1009, 1100, 1101, 1102, 1103, 1109, 1201, 1301, 1399, 1400, 1501, 1599, 1699

Τα δεδομένα είναι ήδη ταξινομημένα

1	29
2	179
3	180
4	201
5	300
6	301
7	304
8	350
9	399
10	401
11	455
12	501
13	503
14	540
15	543
16	549
17	560
18	561
19	562
20	563
21	569
22	570
23	599
24	601
25	603
26	650

27	701
28	703
29	704
30	709
31	713
32	733
33	745
34	801
35	900
36	982
37	983
38	985
39	999
40	1001
41	1002
42	1003
43	1009
44	1100
45	1101
46	1102
47	1103
48	1109
49	1201
50	1301
51	1399
52	1400
53	1501
54	1599
55	1699

Πλήθος δεδομένων: n=55

Τώρα θα δούμε τον καλύτερο τρόπο υπολογισμού της διαμέσου και των Q_1 και Q_3 (όπως αναφέραμε προηγουμένως)

- Όλα τα παρακάτω αναφέρονται σε δεδομένα ταξινομημένα από το μικρότερο προς το μεγαλύτερο

Για να βρούμε τη θέση της **διαμέσου**: (n+1)/2 = (55+1)/2 = 56/2 = 28

- Άρα, η διάμεσος (median) θα είναι το δεδομένο που βρίσκεται στη θέση 28, δηλαδή το 703

(Η διάμεσος δεν είναι απαραίτητη για τον υπολογισμό των φραχτών αλλά την υπολογίσαμε για να δείξουμε τον βελτιωμένο τρόπο υπολογισμού της θέσης της)

Για να βρούμε τη θέση του **Q_1**: (n+1)×1/4 = (n+1)/4 = (55×1)/4 = 56/4 = 14

- Επομένως, το Q_1 θα είναι το δεδομένο που βρίσκεται στη θέση 14, δηλαδή το 540

Για να βρούμε τη θέση του **Q_3**: (n+1)×3/4 = (55×1)×3/4 = 56×3/4 = 42

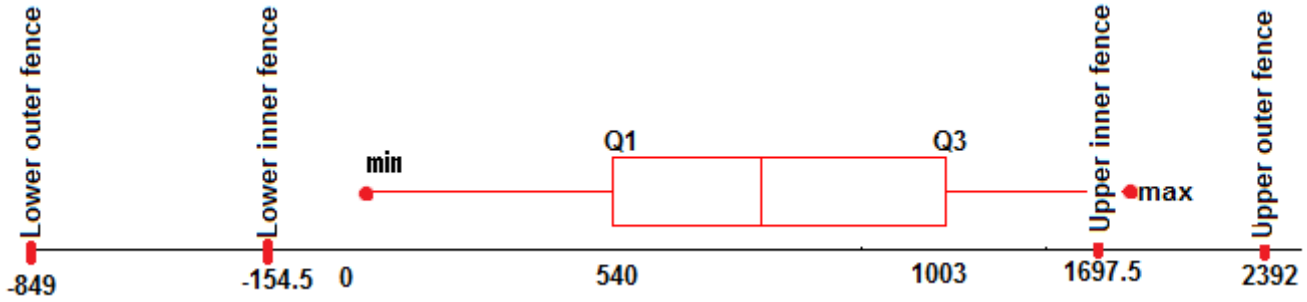
- Επομένως, το Q_3 θα είναι το δεδομένο που βρίσκεται στη θέση 42, δηλαδή το 1003

Ενδοτεταρτημοριακό πλάτος: **IQR** = $Q_3 - Q_1$ = 1003 − 540 = 463

Υπολογισμός φραχτών

- Lower outer fence: $Q_1 - 3\times IQR = 540 - 3\times463 = -849$
- Lower inner fence: $Q_1 - 1.5\times IQR = 540 - 1.5\times463 = -154.5$
- *(Η διάμεσος, που ισούται με 703, βρίσκεται ανάμεσα στο lower inner και το upper inner fence, χωρίς αυτό να παίζει κανένα ρόλο στο χαρακτηρισμό δεδομένων ως outliers)*
- Upper inner fence: $Q_3 + 1.5\times IQR = 1003 + 1.5\times463 = 1697.5$
- Upper outer fence: $Q_3 + 3\times IQR = 1003 + 3\times463 = 2392$

Αυτά όλα μπορούν να απεικονιστούν διαγραμματικά στο παρακάτω box plot



Παρατηρούμε ότι μόνο ένα δεδομένο, το 55ο στη σειρά και μεγαλύτερο με τιμή 1699 είναι εκτός των εσωτερικών φραχτών αλλά εντός των εξωτερικών.

- Το δεδομένο αυτό είναι πιθανό outlier

Θα μάθουμε ακόμα έναν τρόπο εντοπισμού outliers όταν εξετάσουμε τον εμπειρικό κανόνα (empirical rule)

7.5. Μέση τιμή και τυπική απόκλιση

Η **μέση τιμή** (mean ή average, λέγεται και μέσος όρος) ενός συνόλου δεδομένων (δείγματος) υπολογίζεται εάν αθροίσουμε όλα τα δεδομένα και διαιρέσουμε με το πλήθος τους

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Η μέση τιμή ενός δείγματος συμβολίζεται με $\bar{x}$ και αποκαλείται “x-bar”

Παράδειγμα από εδώ: εκπομπές διοξειδίου του άνθρακα (CO_2) για τις ακόλουθες 5 χώρες με τις περισσότερες εκπομπές

ΑΑ	*Χώρα*	*Εκπομπές CO_2 του 2016 (δισεκατομμύρια τόνοι)*
1	Κίνα	10.433
2	ΗΠΑ	5.012
3	Ινδία	2.534
4	Ρωσία	1.662
5	Ιαπωνία	1.24

Ο μέσος όρος των εκπομπών για αυτές τις 5 χώρες είναι

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{10.433 + 5.012 + 2.534 + 1.662 + 1.24}{5} = \frac{20.881}{5} = 4.176$$

Ο μέσος όρος λοιπόν των ετήσιων εκπομπών CO_2 του έτους 2016 για αυτές τις 5 χώρες είναι 4.186 δισεκατομμύρια τόνοι

- Έχει ενδιαφέρον να παρατηρήσουμε ότι η αριθμητική τιμή του μέσου όρου, 4.176, δεν ισούται με κανένα από τα δεδομένα του δείγματος

Η **τυπική απόκλιση** (standard deviation) ενός συνόλου δεδομένων (δείγματος) υπολογίζεται με τον ακόλουθο τύπο

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(\bar{x} - x_i\right)^2}{n-1}} = \sqrt{\frac{\left(\bar{x} - x_1\right)^2 + \left(\bar{x} - x_2\right)^2 + \cdots + \left(\bar{x} - x_n\right)^2}{n-1}}$$

Βλέπετε το $\bar{x}$ στον παραπάνω τύπο;

- Για να υπολογίσουμε την τυπική απόκλιση ενός συνόλου δεδομένων, πρέπει πρώτα να υπολογίσουμε τη μέση τιμή

Είναι επίσης περίεργο που στον παρονομαστή του κλάσματος μέσα στην τετραγωνική ρίζα, έχουμε n−1 και όχι n

- Αυτό θα το εξηγήσουμε στο κοντινό μέλλον

Υπολογισμός μέσης τιμής για τις ετήσιες εκπομπές CO_2 των 5 παραπάνω χωρών

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(\bar{x} - x_i\right)^2}{n-1}}$$

$$= \sqrt{\frac{(4.176-10.433)^2+(4.176-5.012)^2+(4.176-2.534)^2+(4.176-1.662)^2(4.176-1.24)^2}{5-1}}$$

$$= \sqrt{\frac{(-6.257)^2+(-0.836)^2+(1.642)^2+(2.514)^2+(2.936)^2}{4}}$$

$$= \sqrt{\frac{39.15+0.699+2.696+6.32+8.62}{4}}$$

$$= \sqrt{\frac{57.485}{4}} = \sqrt{14.371} = 3.791$$

Η τυπική απόκλιση λοιπόν των ετήσιων εκπομπών CO_2 του έτους 2016 για αυτές τις 5 χώρες είναι 3.791 δισεκατομμύρια τόνοι

- Η τυπική απόκλιση δείχνει (*κατά κάποιο τρόπο*) τη μέση απόσταση των δεδομένων από τη μέση τιμή του δείγματος

Το τετράγωνο της τυπικής απόκλισης λέγεται **διακύμανση** (variance), και για τις 5 παραπάνω χώρες είναι ίσο με

$$v = s^2 = \frac{\sum_{i=1}^{n}\left(\bar{x}-x_i\right)^2}{n-1} = \frac{57.485}{4} = 14.371$$

Αντίστοιχα, μπορούμε να πούμε ότι η τυπική απόκλιση είναι η τετραγωνική ρίζα της διακύμανσης

$$s = \sqrt{v} = \sqrt{\frac{\sum_{i=1}^{n}\left(\bar{x}-x_i\right)^2}{n-1}}$$

Η διακύμανση υπολογίζεται όπως η τυπική απόκλιση χωρίς όμως την τετραγωνική ρίζα

- *Προσοχή όμως: οι μονάδες της διασποράς είναι το τετράγωνο των μονάδων των δεδομένων!*
    - Επομένως, η διασπορά των ετήσιων εκπομπών CO_2 των 5 παραπάνω χωρών είναι 14.371 (δισεκατομμύρια τόνοι)²
        - Οι μονάδες της διασποράς τις περισσότερες φορές θα είναι ακαταλαβίστικες, ενώ της τυπικής απόκλισης θα είναι ίδιες με αυτές των αρχικών δεδομένων
        - Όπως θα δούμε αργότερα, η διασπορά είναι χρήσιμη σε μια συγκεκριμένη περίπτωση, κατά κανόνα όμως η τυπική απόκλιση είναι πολύ πιο χρήσιμη (και εύκολη να την καταλάβουμε)

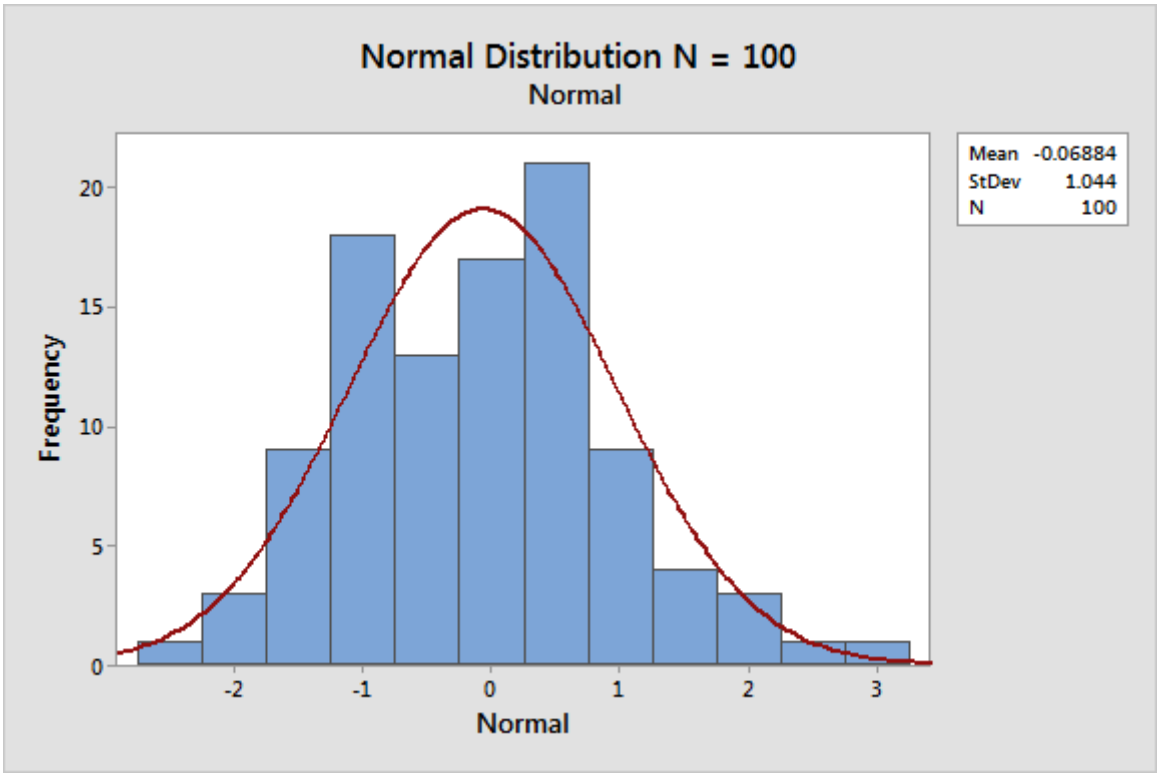
Άσκηση για να εμπεδώσουμε την έννοια της τυπικής απόκλισης — Διαλέξτε 4 ακέραιους αριθμούς από το 1 μέχρι και το 9 – π.χ. (1, 3, 5, 5) δηλαδή μπορεί

κάποιοι ή και όλοι από αυτούς τους 4 αριθμούς να είναι ίδιοι – έτσι ώστε η τυπική τους απόκλισή τους να είναι

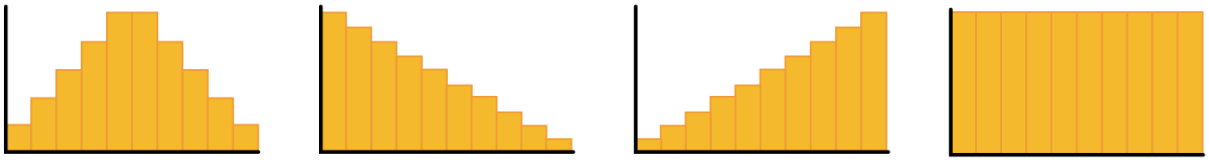
1. Ελάχιστη
    - Αρκεί να είναι οι 4 αριθμοί ίδιοι, τότε η τυπική τους απόκλιση θα είναι μηδέν!
    - Πόσες τέτοιες λύσεις υπάρχουν;
        i. Συνολικά εννέα: (1, 1, 1, 1), (2, 2, 2, 2) … (9, 9, 9, 9)
2. Μέγιστη
    - (1, 1, 9, 9)
    - Πόσες τέτοιες λύσεις υπάρχουν;
        i. Μόνο μία, η παραπάνω

7.6. Συμμετρία και στρέβλωση

Μια εικόνα της κατανομής ενός συνόλου δεδομένων δίνεται από το ιστόγραμμα



Symmetric (normal) vs skewed and uniform distriutions



Normal distribution
(unimodal, symmetric,
the "bell curve")

Right-skewed distribution
(Positively-skewed)

Left-skewed distribution
(Negatively-skewed)

Uniform distribution
(equal spread,
no peaks)

Unimodal vs bimodal distributions

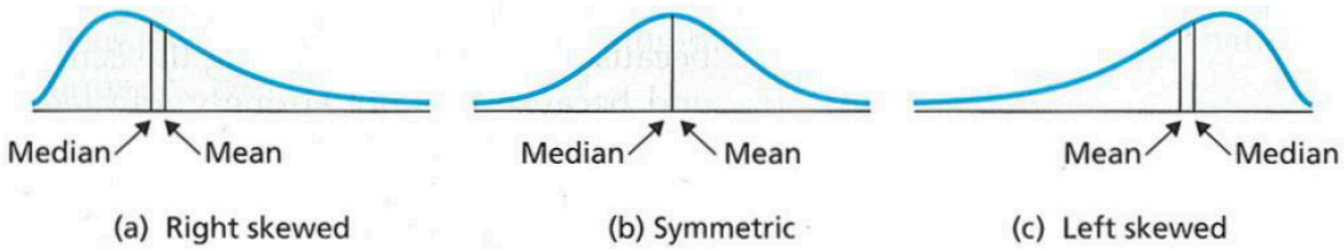
Normal distribution
(unimodal, symmetric,
the "bell curve")

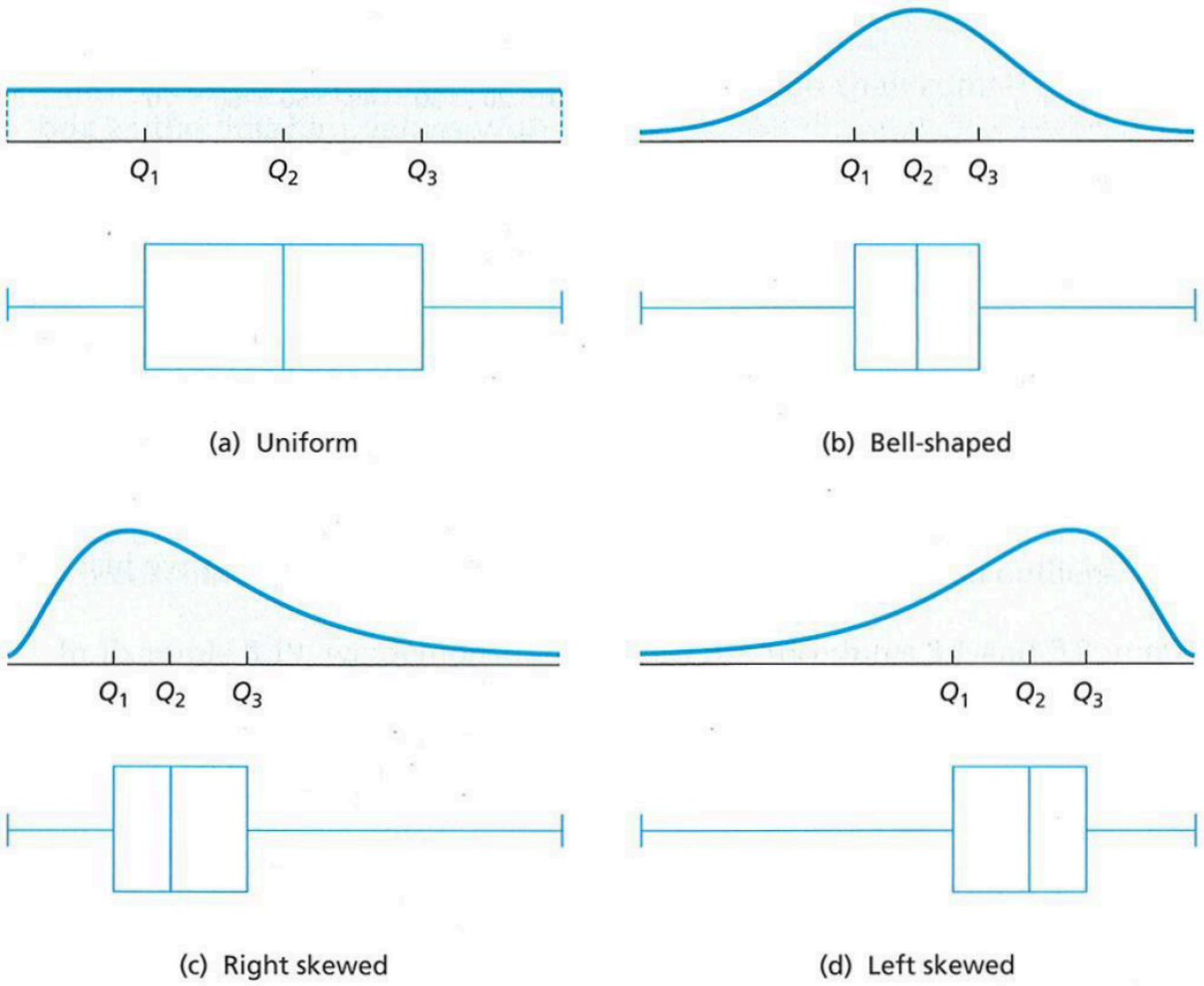
Symmetric bimodal distribution
(two modes)

Non-symmetric bimodal distribution
(two modes)

Στρέβλωση (skewness)

- Συμμετρικές κατανομές: μηδενική στρέβλωση
- Όταν το σουβλί είναι στα αριστερά: αρνητική στρέβλωση
- Όταν το σουβλί είναι στα δεξιά: θετική στρέβλωση

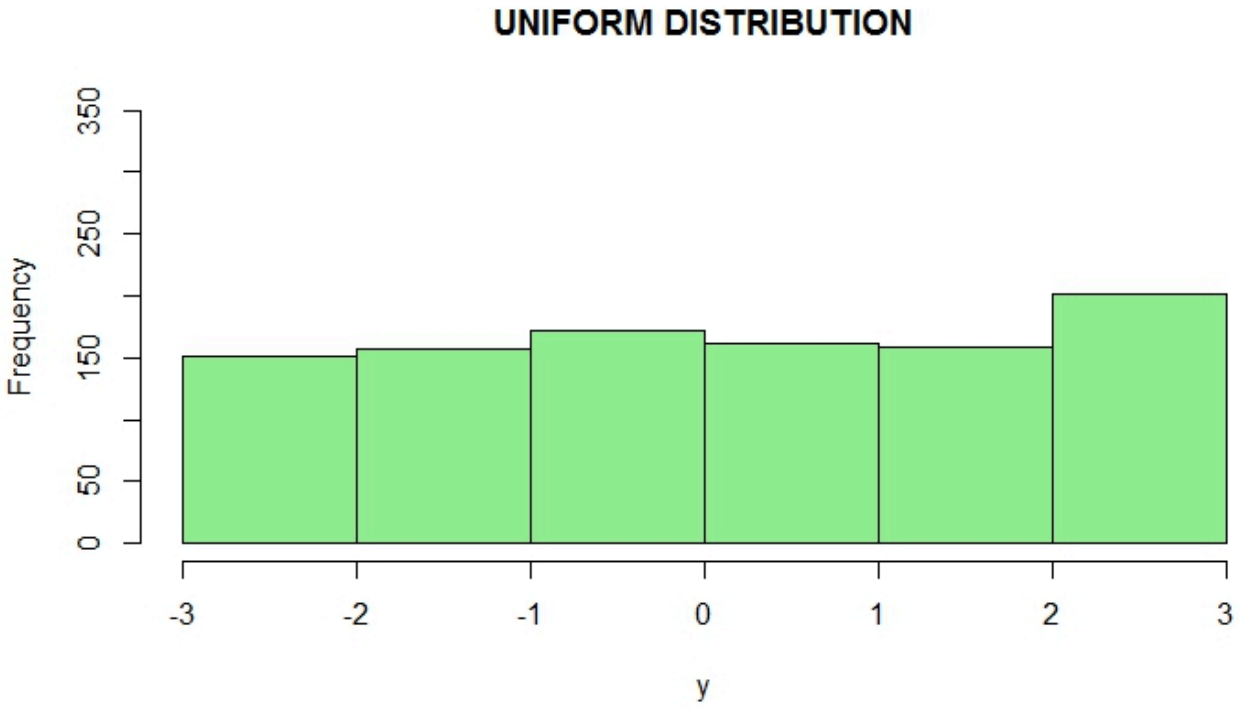


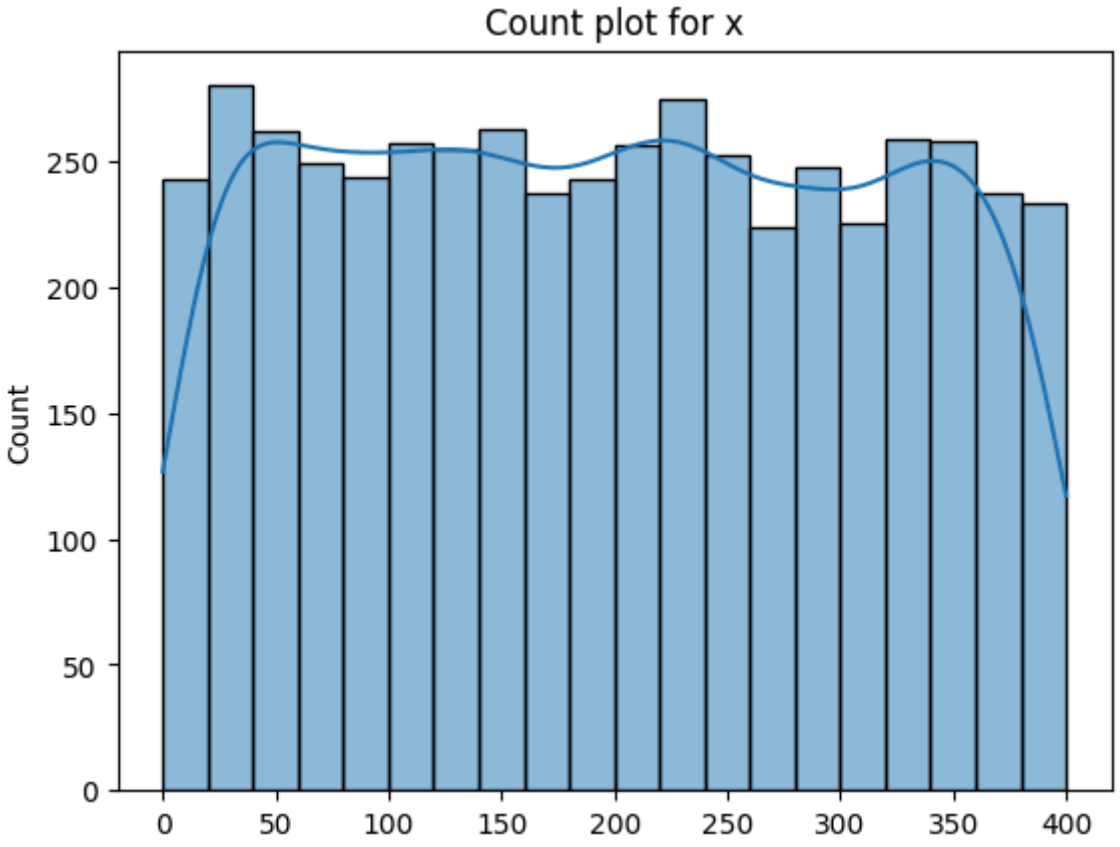


8. Κατανομές πιθανοτήτων

απολύτως ελάχιστες γνώσεις πιθανοτήτων

8.1. Ομοιόμορφη κατανομή





Uniform Distribution Formula

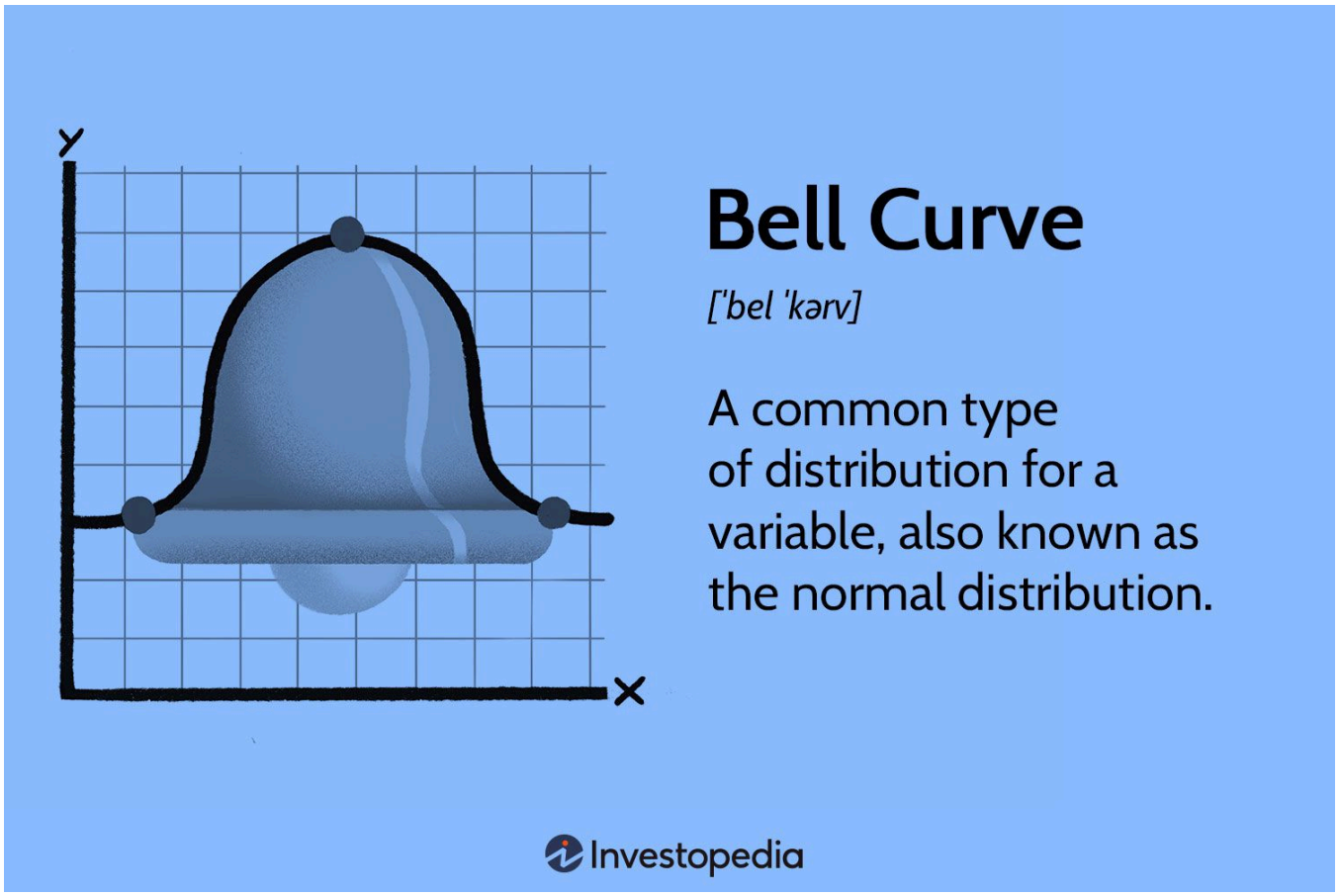
$$F(x) = \frac{1}{(b - a)} \qquad \text{Mean} = \frac{(a + b)}{2}$$

$$\sigma = \sqrt{\frac{(b - a)^2}{12}}$$

8.2. Κανονική κατανομή

Κανονική κατανομή (normal distribution) – αναφέρεται και ως κανονική καμπύλη (normal curve)

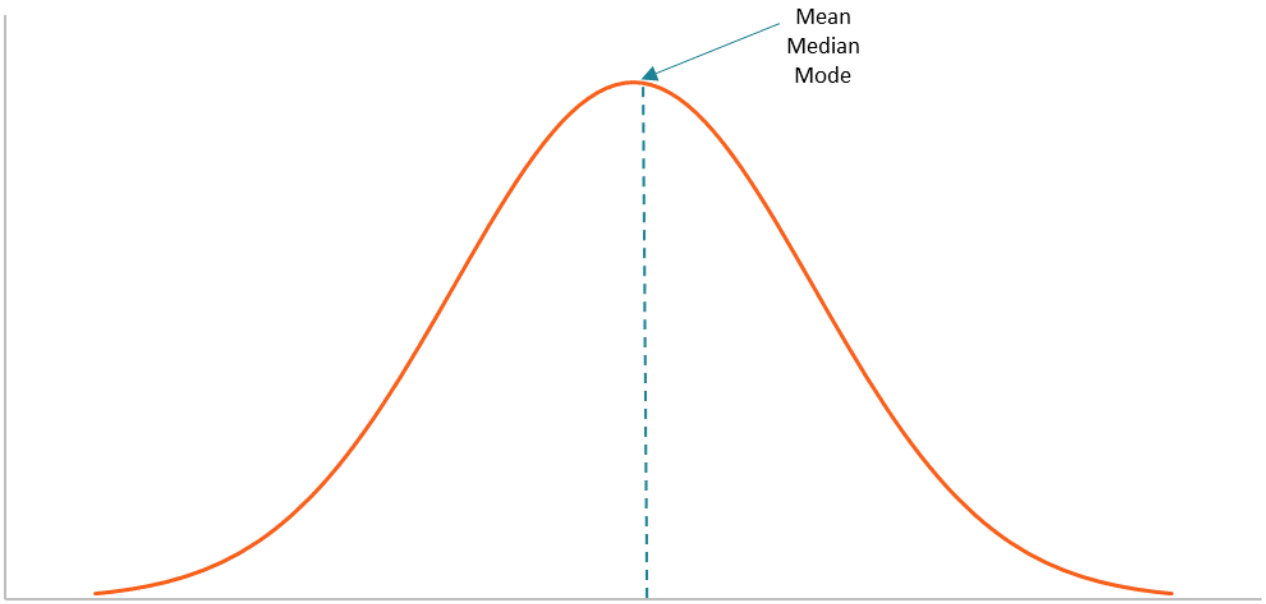
- Καμπύλη σαν καμπάνα (bell curve)



Τυπικά χαρακτηριστικά της κανονικής καμπύλης

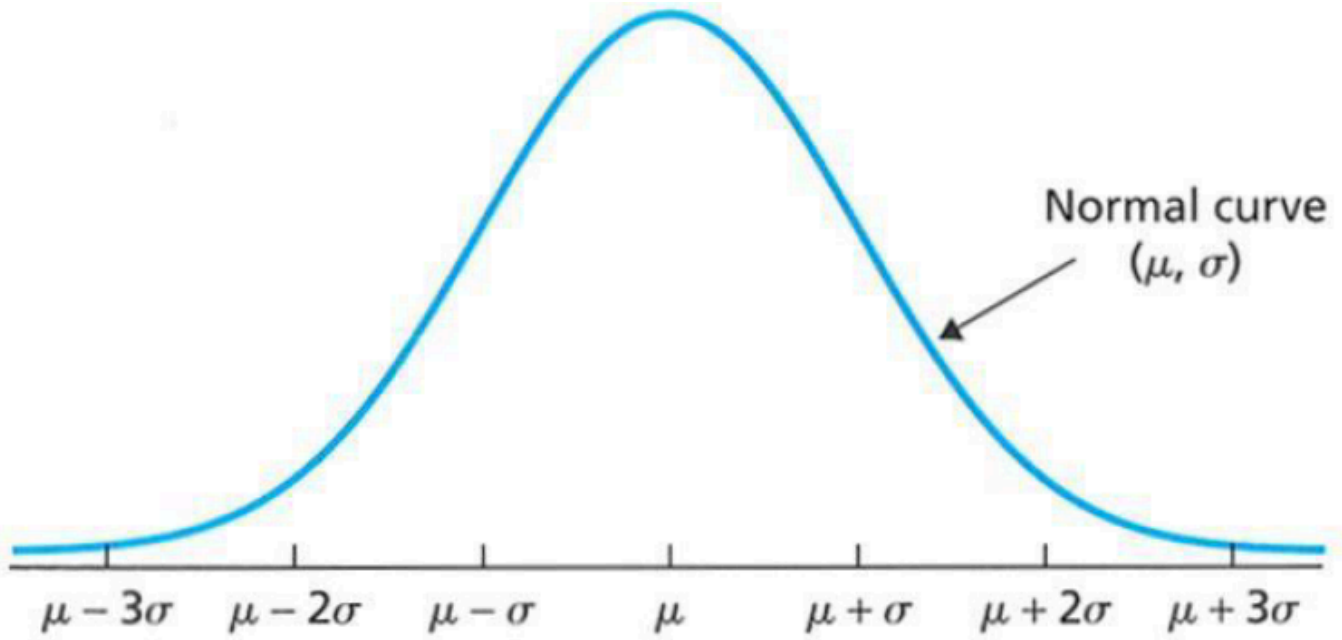
- Συμμετρική
    - Μηδενική στρέβλωση (skewness)

- Ελάχιστο = μείον άπειρο (−∞)
- Μέγιστο = συν άπειρο (+∞)

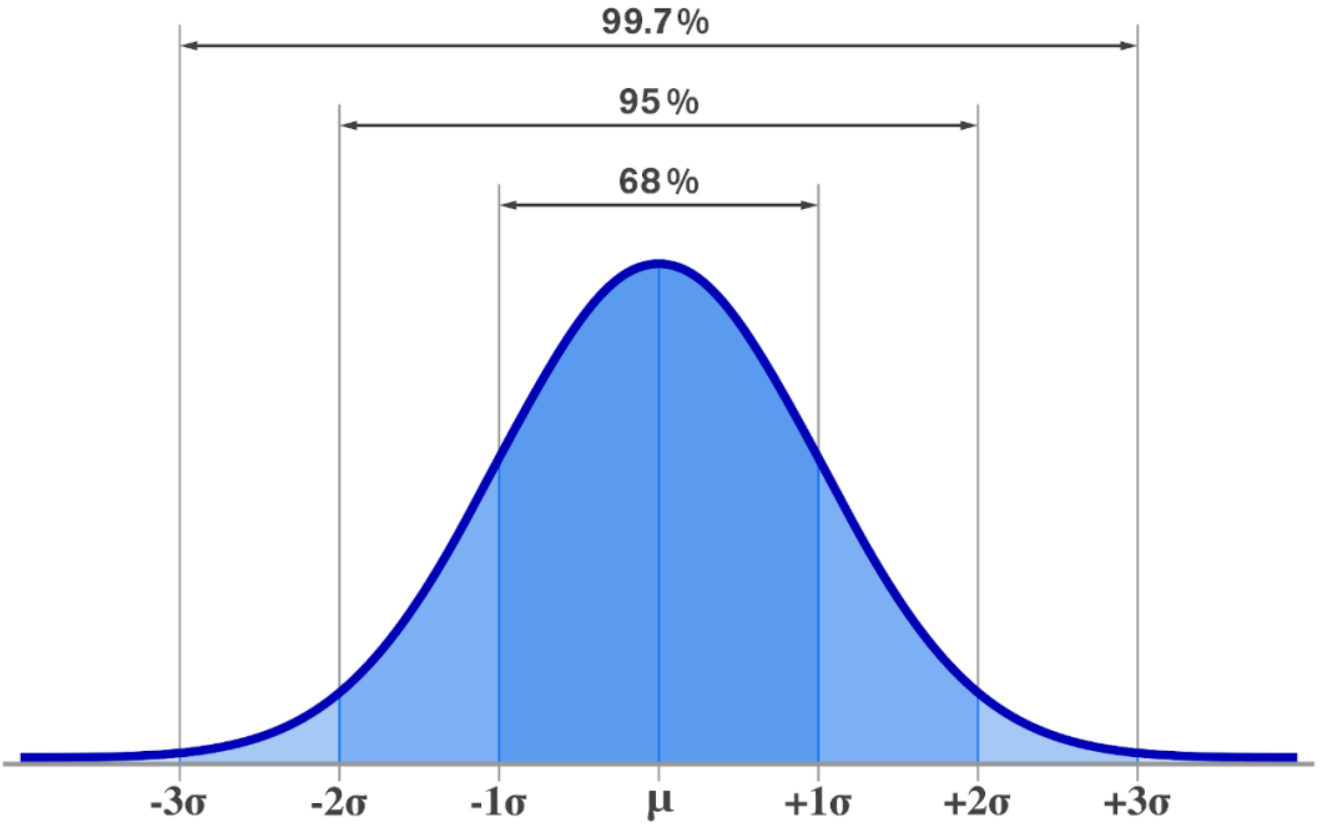


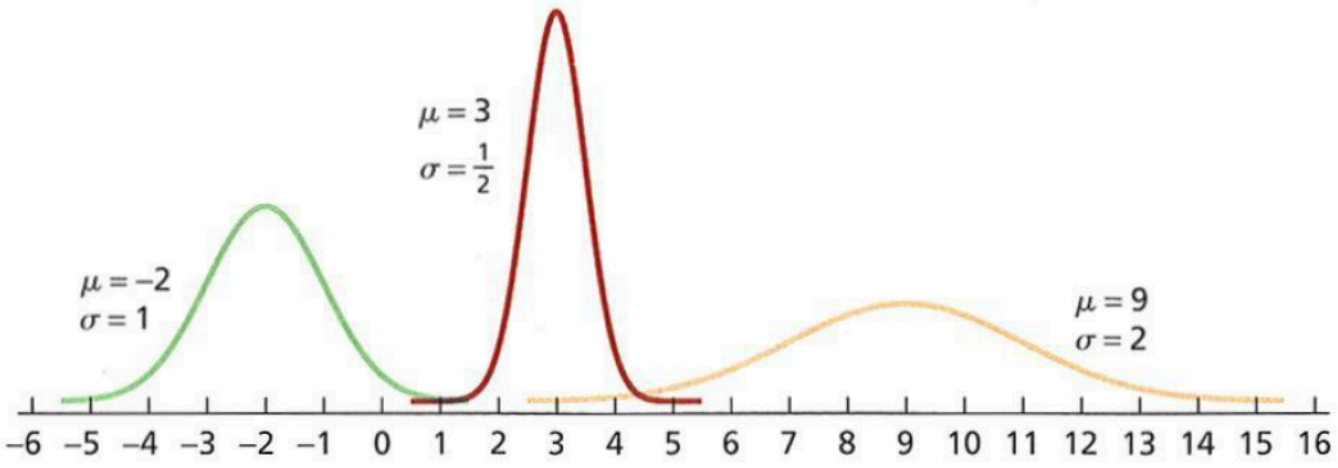
Μια κανονική κατανομή καθορίζεται πλήρως από τη μέση τιμή της, μ, και στην τυπική της απόκλιση, σ

- Και συμβολίζεται N(μ, σ)



Η κανονική κατανομή έχει συγκεκριμένα εμβαδά ανάμεσα στα παρακάτω διαστήματα (που, όμως θα δούμε σε λίγο, αναφέρονται στον εμπειρικό κανόνα)

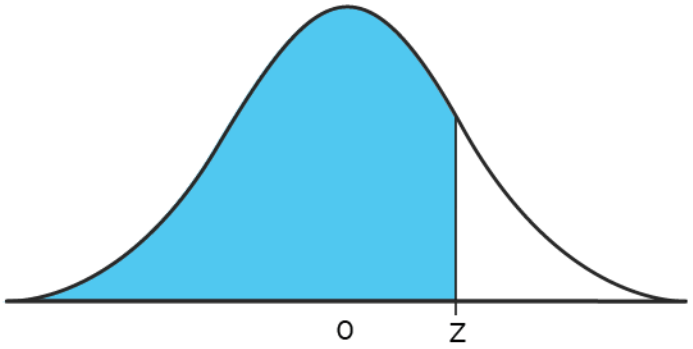




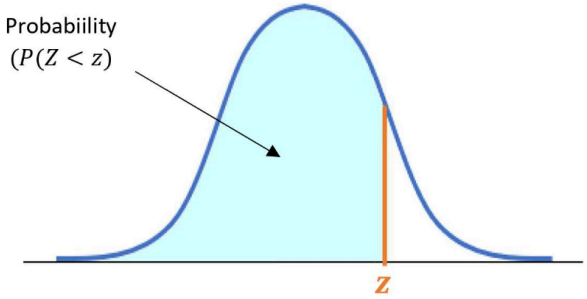
8.2.1. Πίνακες κανονικής κατανομής

τυποποιημένη ή μοναδιαία κανονική κατανομή

Table of Standard Normal Probabilities for Positve Z-scores



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



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	0.5754
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	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	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	0.7996	0.8023	0.8051	0.8079	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	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	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	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	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	0.9865	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	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	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	0.9996	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	0.9998	0.9998

z-scores

raw score

mean

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

standard deviation

Solving for Z Scores

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

Example: $X = 120$, $\mu = 100$, $\sigma = 10$

$$z = \frac{120 - 100}{10} = \frac{20}{10} = +2$$

Interpretation: A score of 120 is two standard deviations above the mean.

Finding a z-score: the formula

$$z = \frac{x - \mu}{\sigma}$$ (where x is the data value)

A set of Economics Final Exam Grades are normally distributed with a mean of 65 and a standard deviation of 12.

Find a z-score for a grade of 70.

$$z = \frac{70 - 65}{12} = 0.42$$

Find a z-score for a grade of 62.

$$z = \frac{62 - 65}{12} = -0.25$$

Find a z-score for a grade of 49.

$$z = \frac{49 - 65}{12} = -1.33$$

My time to run 200m is 28s. The mean time for this race is 31s and the standard deviation is 1.5s.

My time to run 500m is 132s. The mean time for this race is 125s and the standard deviation is 8.2s.

Assuming a normal distribution, in which race did I perform the best compared to the other participants?

200m

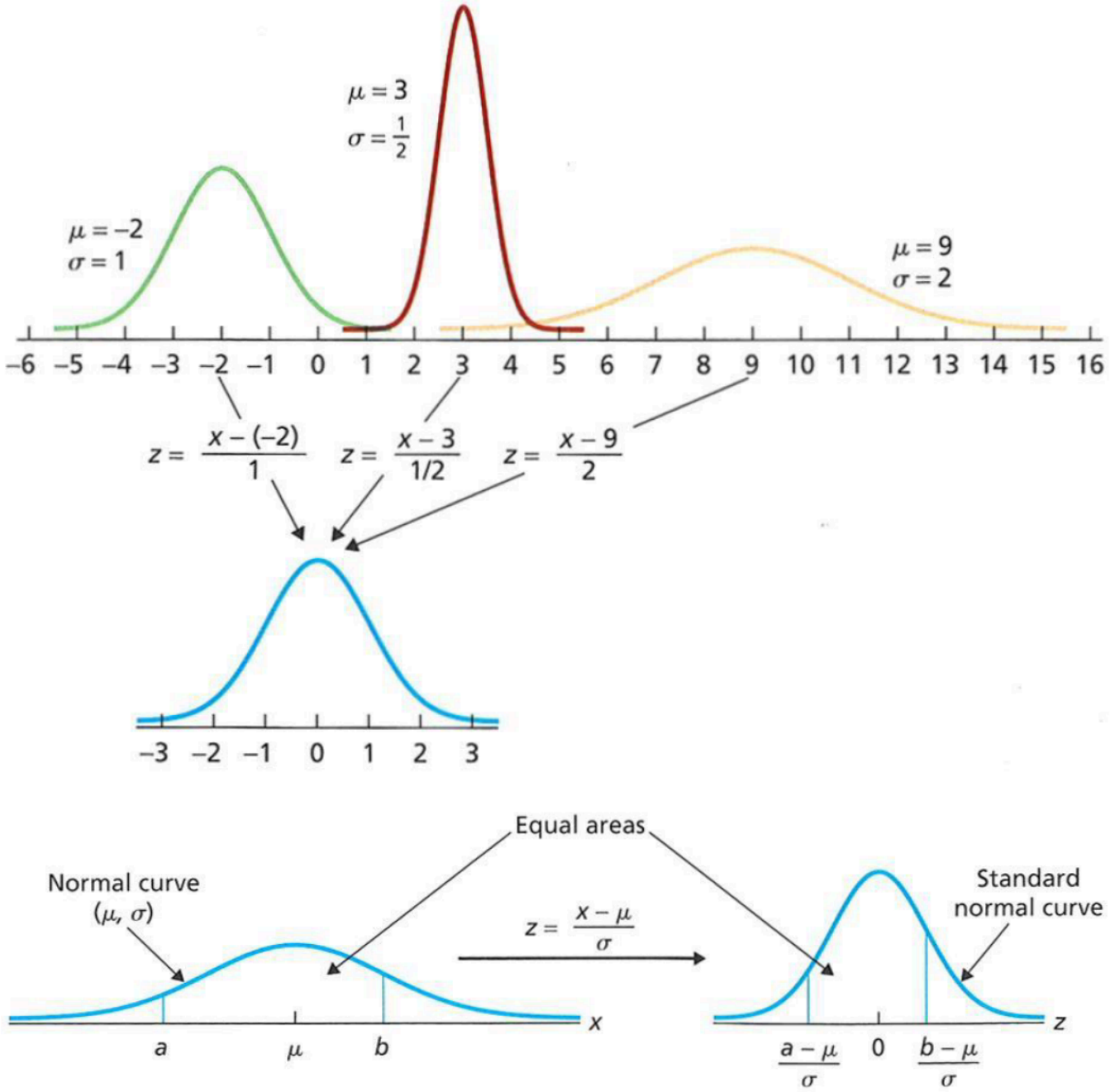
$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{28 - 31}{1.5} = -2$$

500m

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

$$z = \frac{132 - 125}{8.2} = 0.854$$



Αντίστροφοι υπολογισμοί

Determining a Raw Score (X) from a Z-Score: Example #2

$z = -1, \mu = 50, \sigma = 10$; Solve for X

(*Tip*: Since the z is negative the X value must be below the mean.)

$$X = \mu + z\sigma$$
$$X = 50 + (-1)(10)$$
$$X = 50 - 10$$
$$X = 40$$

Tip: If the standard deviation (SD) is 10 and a score is 1 SD below the mean ($z = -1$), then the score is (1 x 10) points, or 10 points below the mean of 50; that is, $X = 40$.

8.2.2. Εμπειρικός κανόνας

(1) Διάστημα (μ−σ, μ+σ)

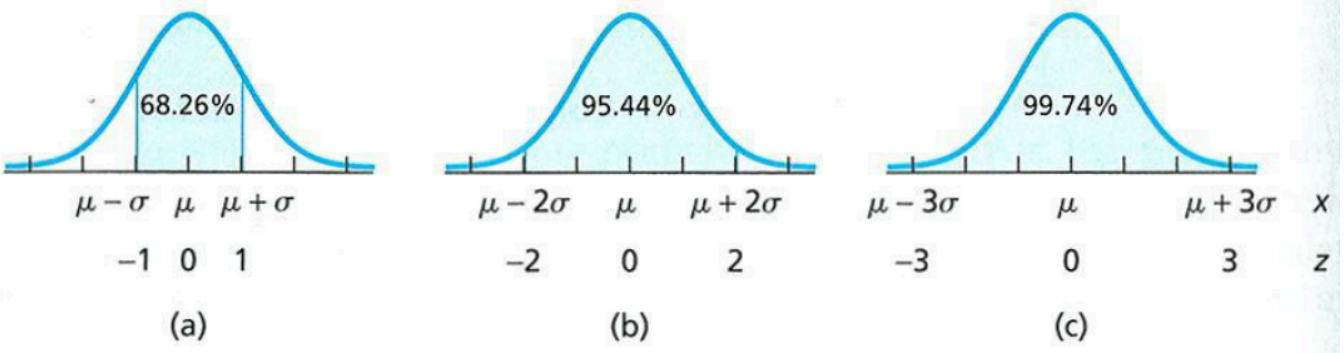
- 68.26%
- (περίπου δύο τρίτα)

(2) Διάστημα (μ−2×σ, μ+2×σ)

- 95.44%
- (περίπου 95%)

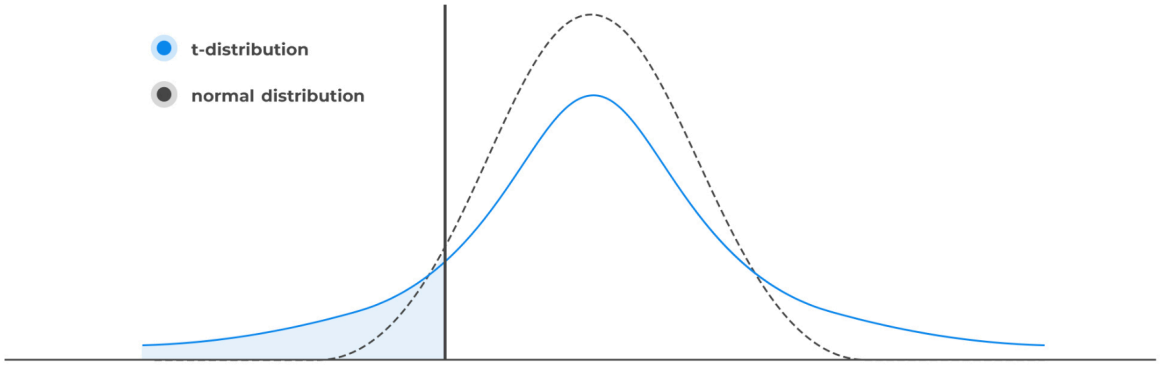
(3) Διάστημα (μ−3×σ, μ+3×σ)

- 99.74%
- (περίπου όλα)



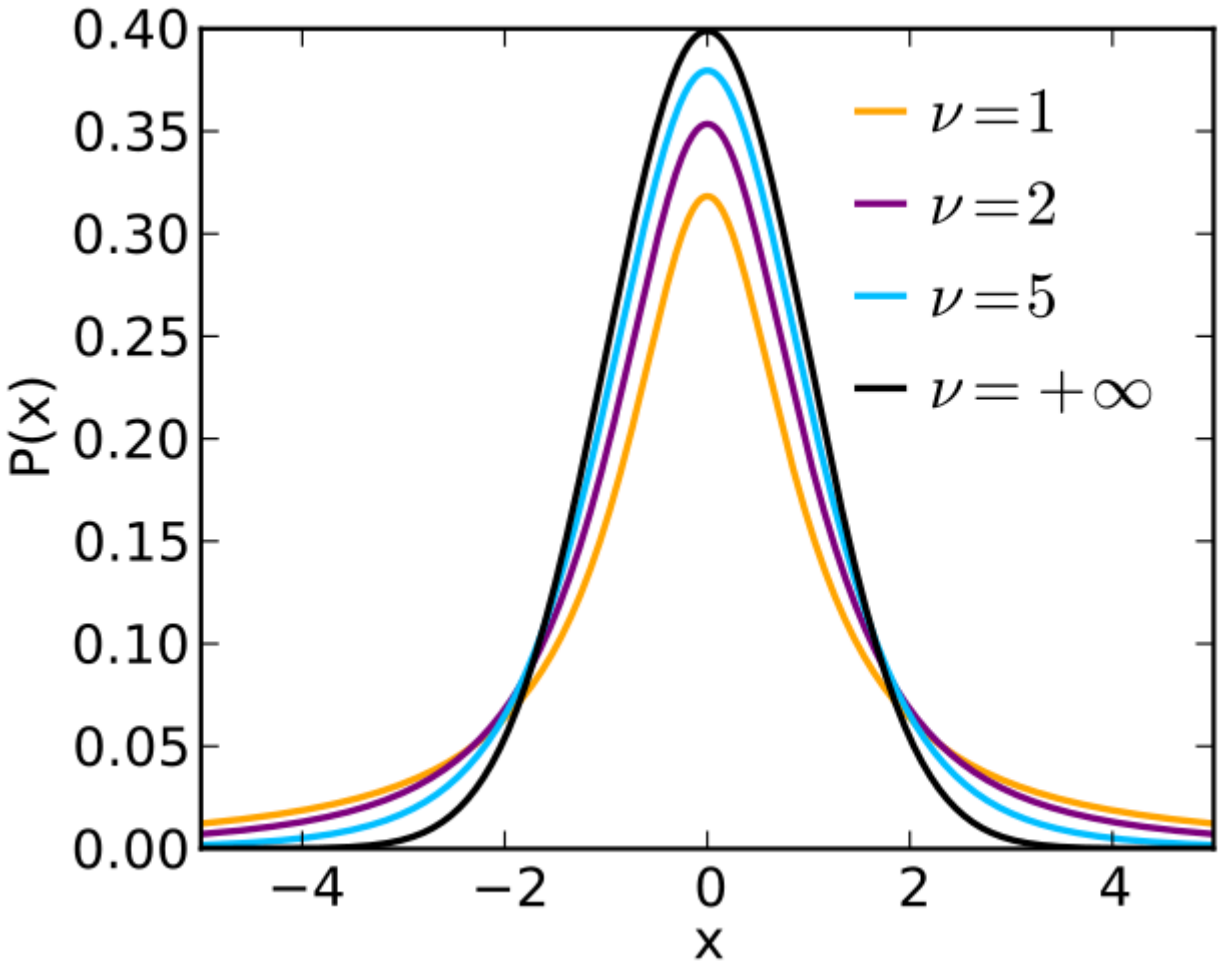
8.3. Κατανομή t (του σπουδαστή)

βαθμοί ελευθερίας (degrees of freedom)



Κανονική κατανομή = κατανομή t με άπειρους (∞) βαθμούς ελευθερίας

- Πάνω από 30 βαθμούς ελευθερίας, η κατανομή t είναι σχεδόν ίδια με την κανονική κατανομή
    - *(Θα δούμε τη σημασία αυτού όταν εξετάσουμε τους στατιστικούς ελέγχους)*



8.4. Κατανομή F

πηλίκο δύο κατανομών t

βαθμοί ελευθερίας (degrees of freedom) του αριθμητή (numerator)

βαθμοί ελευθερίας (degrees of freedom) του παρονομαστή (denominator)

8.5. Κατανομή χ^2

πίνακες κατανομής χ^2

8.6. Κατανομή βήτα

ελάχιστο, κορυφή, μέγιστο

χρήση στον χρονικό και κοστολογικό προγραμματισμό έργων (project management)

9. Δειγματοληπτική κατανομή

από την Περιγραφική Στατιστική περνάμε στην Επαγωγική Στατιστική

10. Πίνακες αληθείας

μηδενική και εναλλακτική υπόθεση

σφάλματα τύπου I και II

πιθανότητα α και β

επίπεδο εμπιστοσύνης (confidence level)

11. Διαστήματα εμπιστοσύνης

της μέσης τιμής

κατά την κατανομή z

κατά την κατανομή t

12.Έλεγχοι υποθέσεων

επίπεδο εμπιστοσύνης

probability p

12.1.Έλεγχος z

της μέσης τιμής

z-test

χρήση πινάκων τυποποιημένης κανονικής κατανομής

12.2.Έλεγχος t

της μέσης τιμής

t-test

χρήση πινάκων κατανομής t

12.3. Έλεγχος χ^2

έλεγχος προσαρμογής σε θεωρητική κατανομή

χρήση πινάκων κατανομής χ^2

έλεγχος σημαντικότητας πινάκων (contingency tables)