

Σημειώσεις Ποσοτικών Μεθόδων



Ιωάννης Αθ. Παραβάντης
Καθηγητής

Εαρινό εξάμηνο 2026

Διδάσκων

Επικοινωνία με διδάσκοντα

- Γιάννης Παραβάντης
 - “Κύριε Παραβάντη” ή “Δάσκαλε”
- Πολιτικός Μηχανικός – Συγκοινωνιολόγος (ΕΜΠ), MSc Transportation & PhD Environmental Health ([Northwestern University](https://www.northwestern.edu))
- 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:
 - Δευτέρα 13:00 με 14:00
 - Πέμπτη 14:00 με 15:00
- Απουσία διδάσκοντα σε ΗΠΑ (με μεταπτυχιακούς)
 - Εβδομάδα 20-26 Απριλίου 2026 δεν θα γίνουν μαθήματα Ποσοτικών Μεθόδων

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

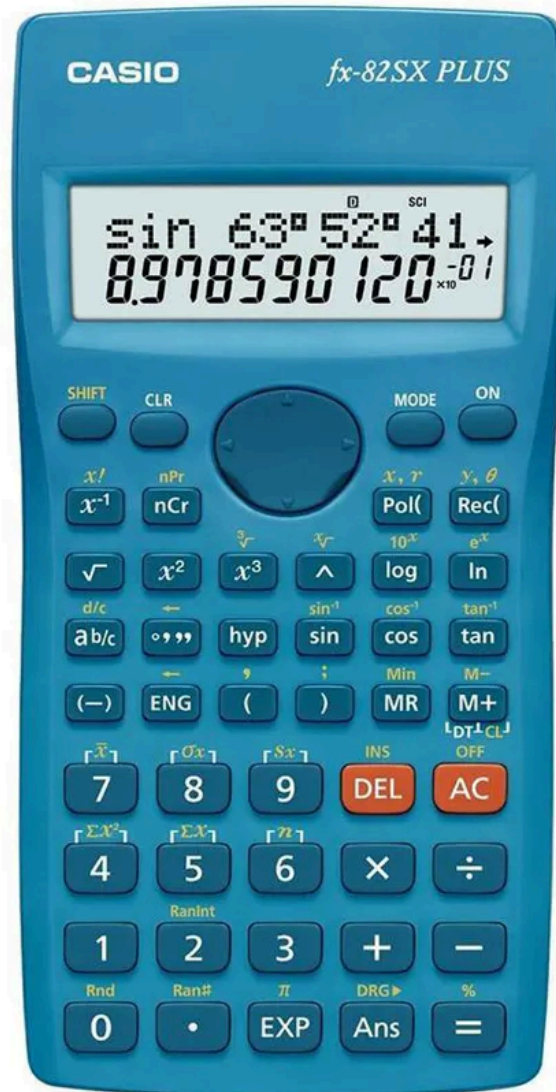
- <https://eclass.unipi.gr/courses/DES103/>
-

1.3. Αξιολόγηση

Μορφές αξιολόγησης

- **Πρόοδοι** (δύο, multiple choice, με προειδοποίηση στο προηγούμενο μάθημα) με κλειστό το εκπαιδευτικό υλικό: **30%** του τελικού βαθμού
 - Αν κάποια πρόοδος γίνει μέσω του Internet, για να μπορέσετε να συμμετάσχετε 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)
- Το κύριο πρόγραμμα εισαγωγής και διαχείρισης δεδομένων, που θα χρησιμοποιήσουμε στο μάθημα

JASP

- Εξαιρετικό δωρεάν πρόγραμμα, κάνει τη χρήση του [R](#) πολύ εύκολη
- Το κύριο πρόγραμμα στατιστικής ανάλυσης που θα χρησιμοποιήσουμε στο μάθημα

jamovi

- Εξαιρετικό δωρεάν πρόγραμμα, ομοίως κάνει τη χρήση του [R](#) πολύ εύκολη
 - Πιο χρήσιμο για συγκεκριμένες στατιστικές αναλύσεις, ως συμπληρωματικό του jamovi
-

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, JASP)

Σημαντικό να γίνεται το μάθημα κατανοητό κατά τη διάρκεια της παράδοσης

- Αλλά όταν μπορούμε στην **Επαγωγική Στατιστική**, εάν παράλληλα με την παρακολούθηση δεν διαβάσετε, 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 διαβάσετε

- 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

Observation #2

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

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

- Διάγραμμα κουτί (boxplot), ιστόγραμμα (histogram), ραβδόγραμμα (barchart)
- Διαγράμματα τύπου πίτας (pie charts) δεν χρησιμοποιούνται πολύ στη στατιστική
- Σε κάθε περίπτωση, αποφεύγουμε εντελώς τα τρισδιάστατα διαγράμματα (αυτό θα το αναφέρουμε ξανά σύντομα)

Monthly Deaths from Lung Diseases in the UK

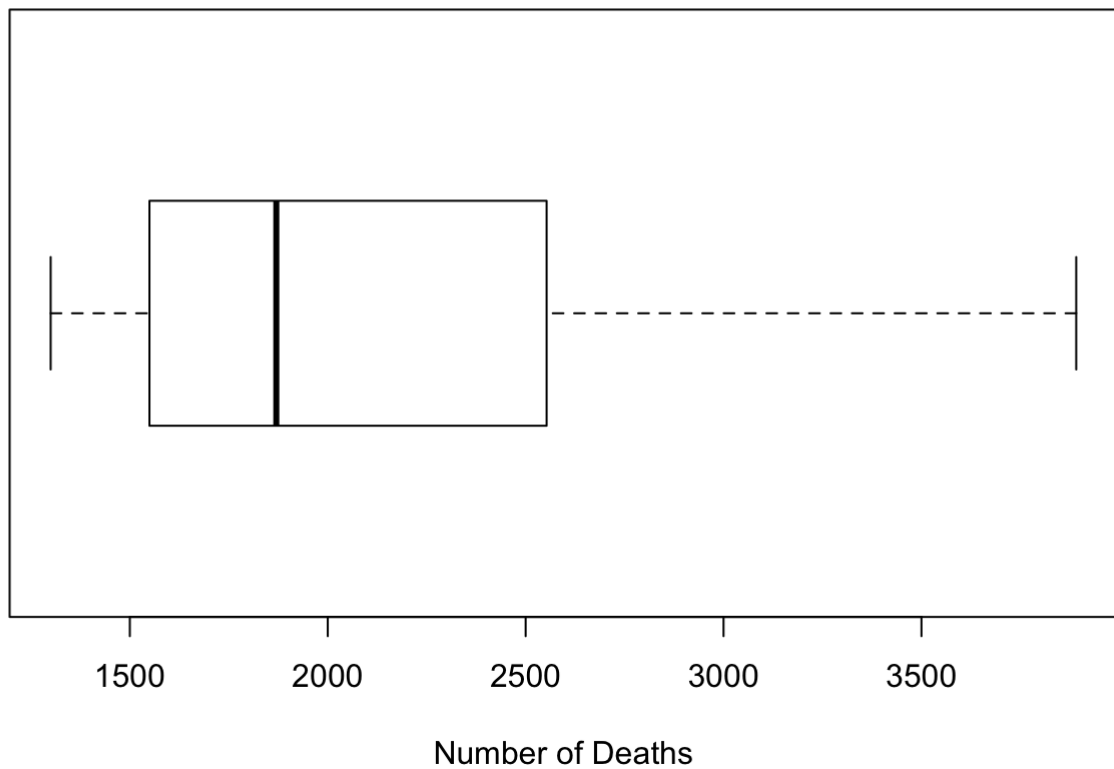
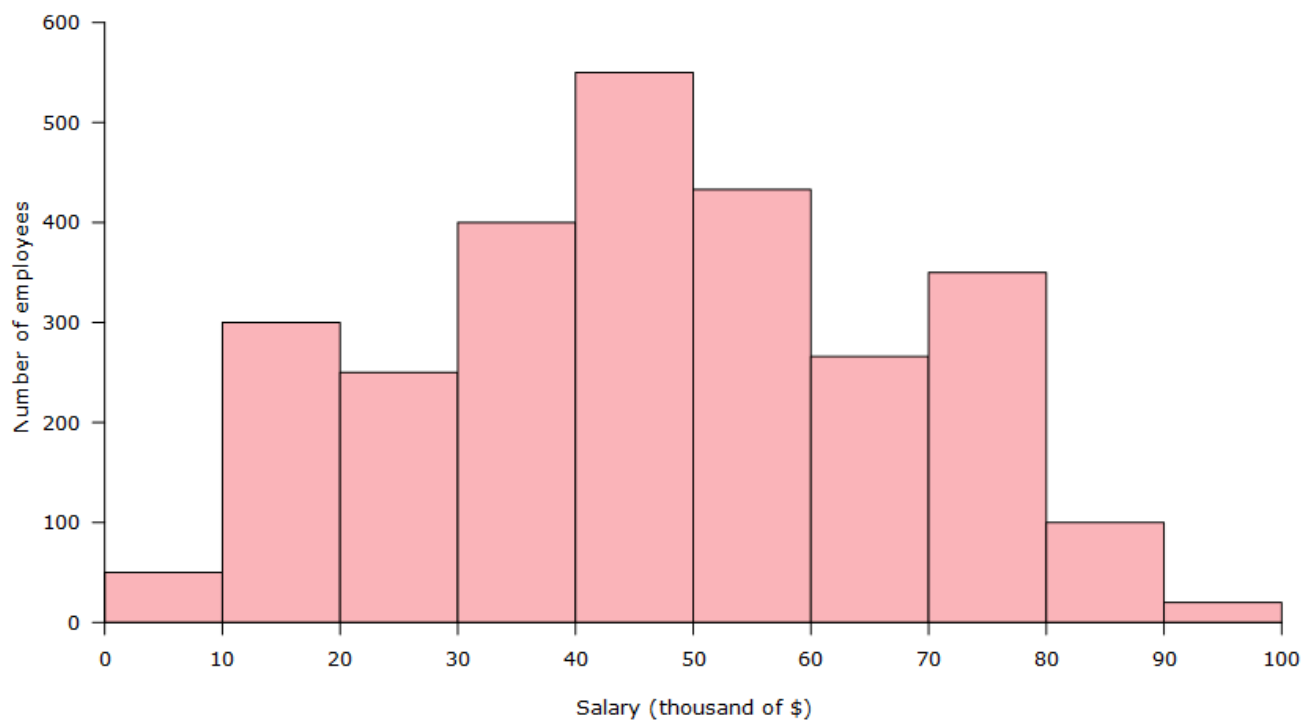
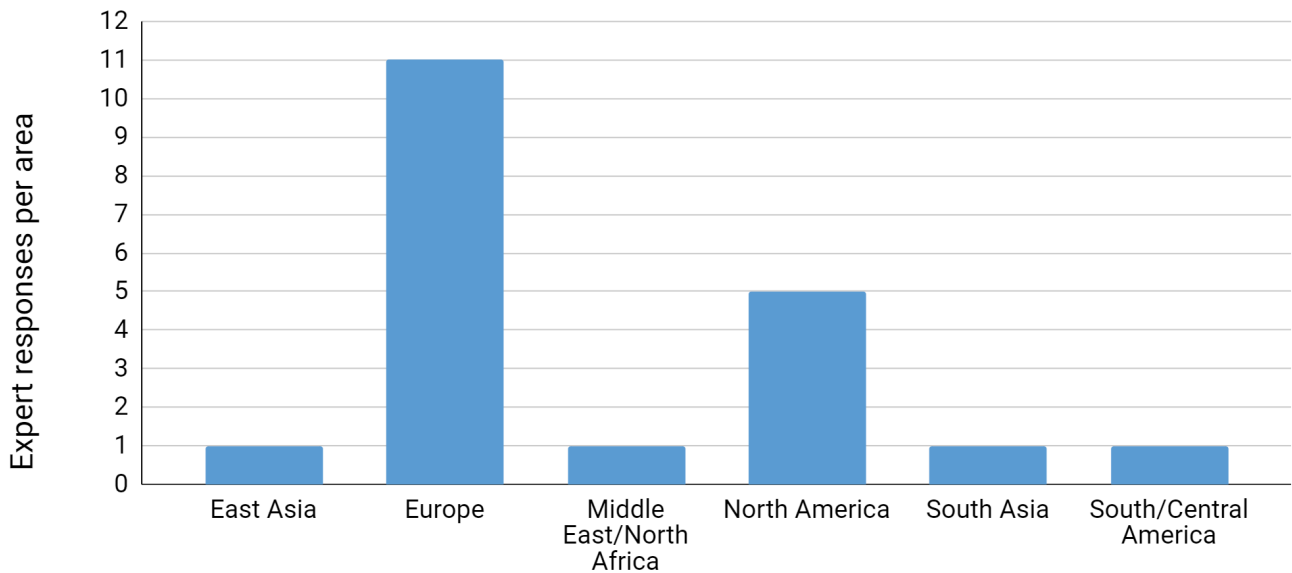


Chart 5.7.1
Distribution of salaries of the employees of ABC Corporation





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



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

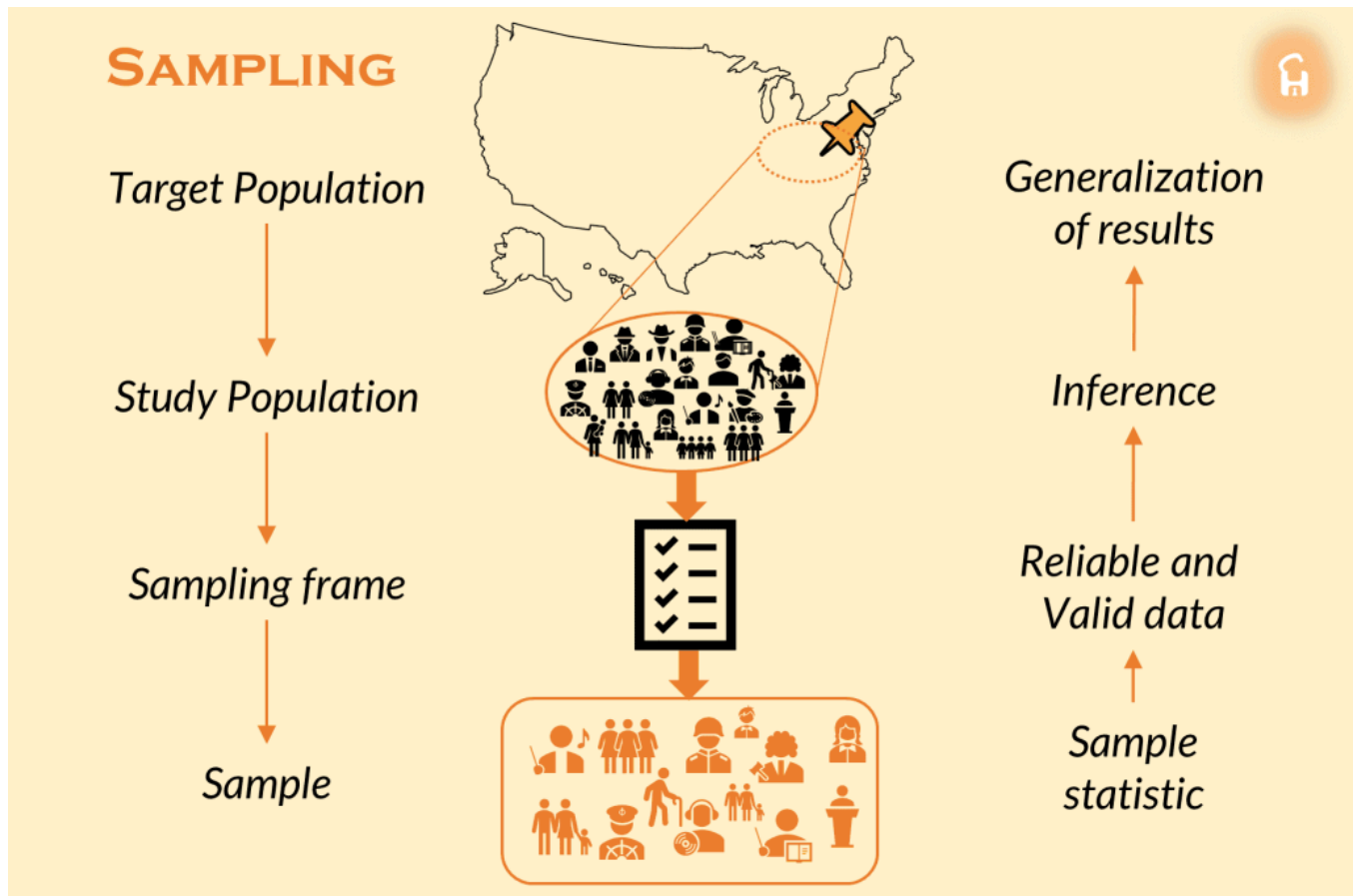
- Κυρίως στο ιστόγραμμα (histogram)

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)

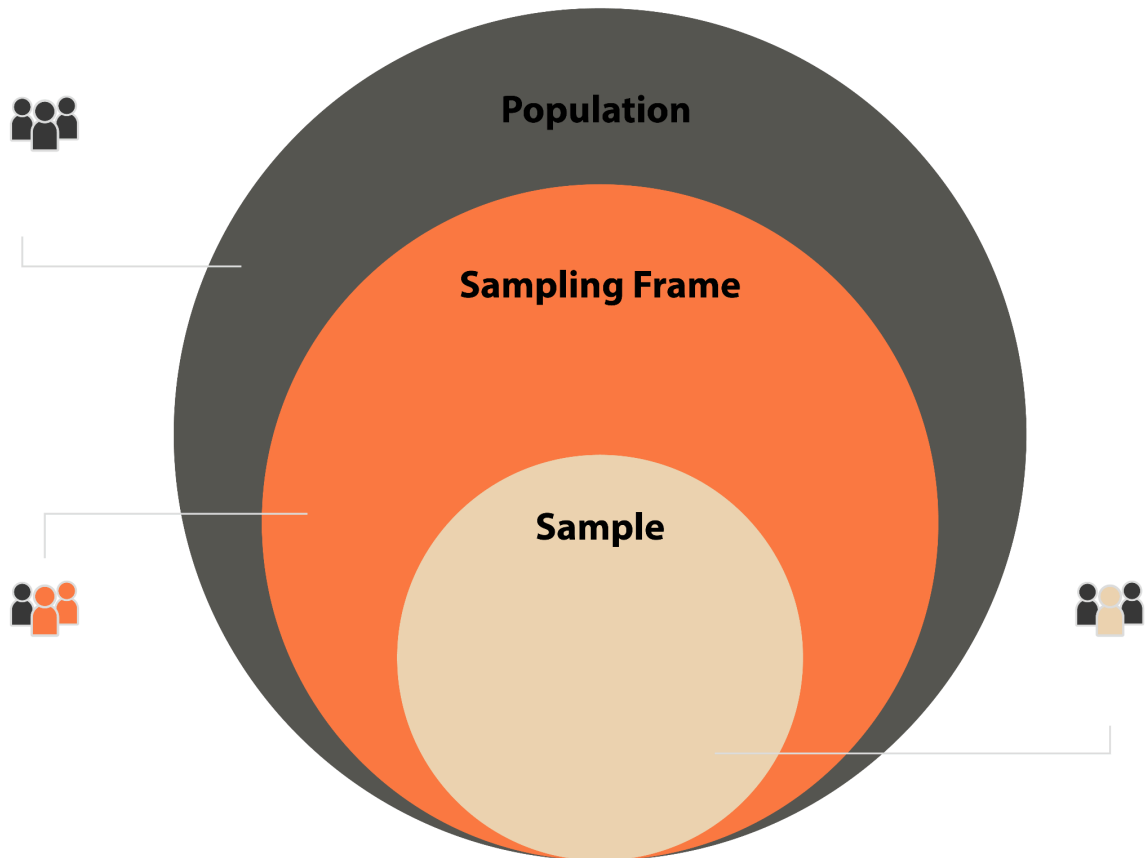


3. Πληθυσμός και δείγμα

Πληθυσμός (population)

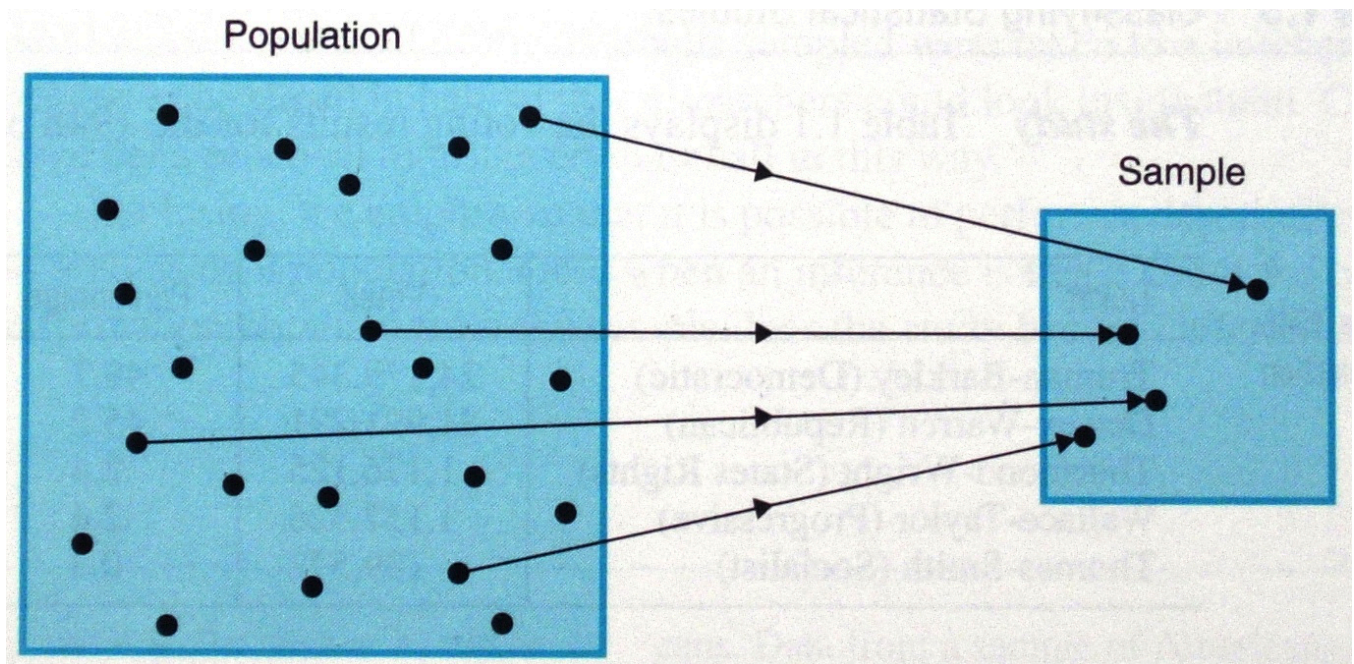
Δειγματοληπτικό πλαίσιο (sampling frame)

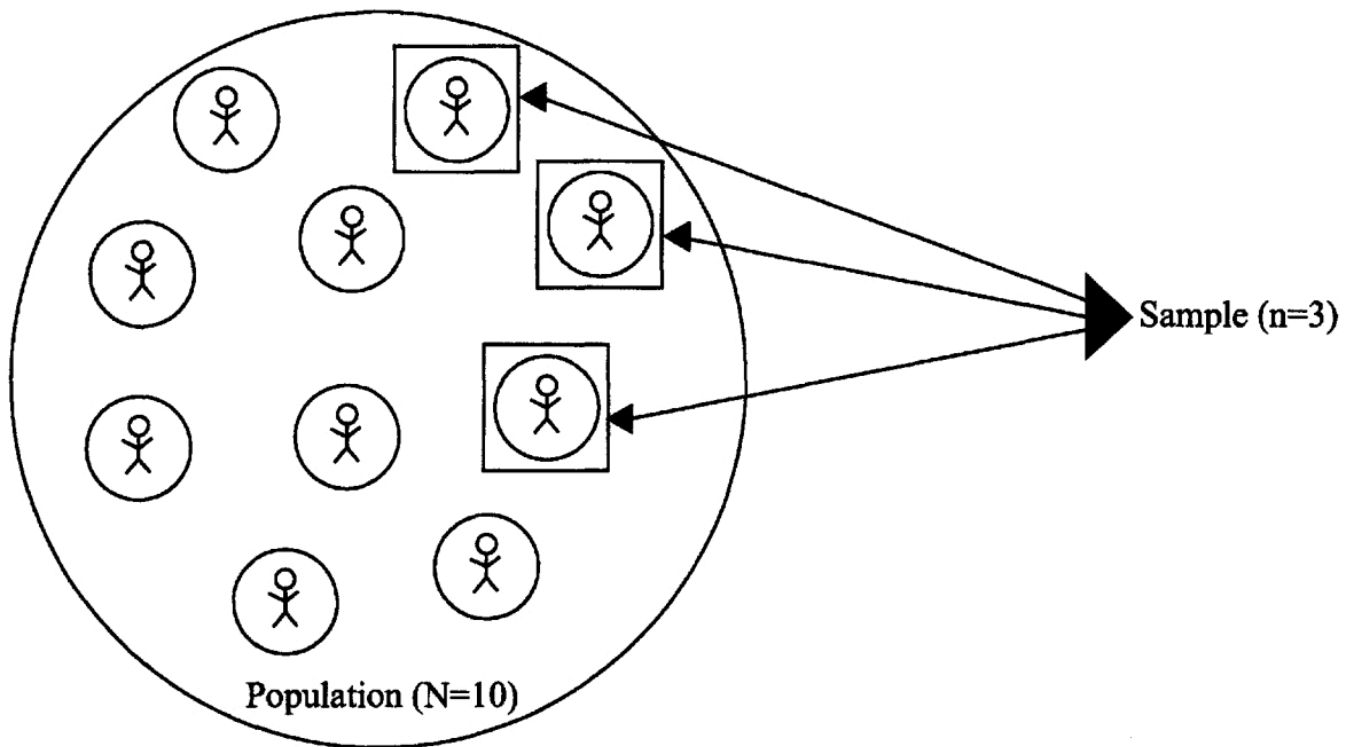
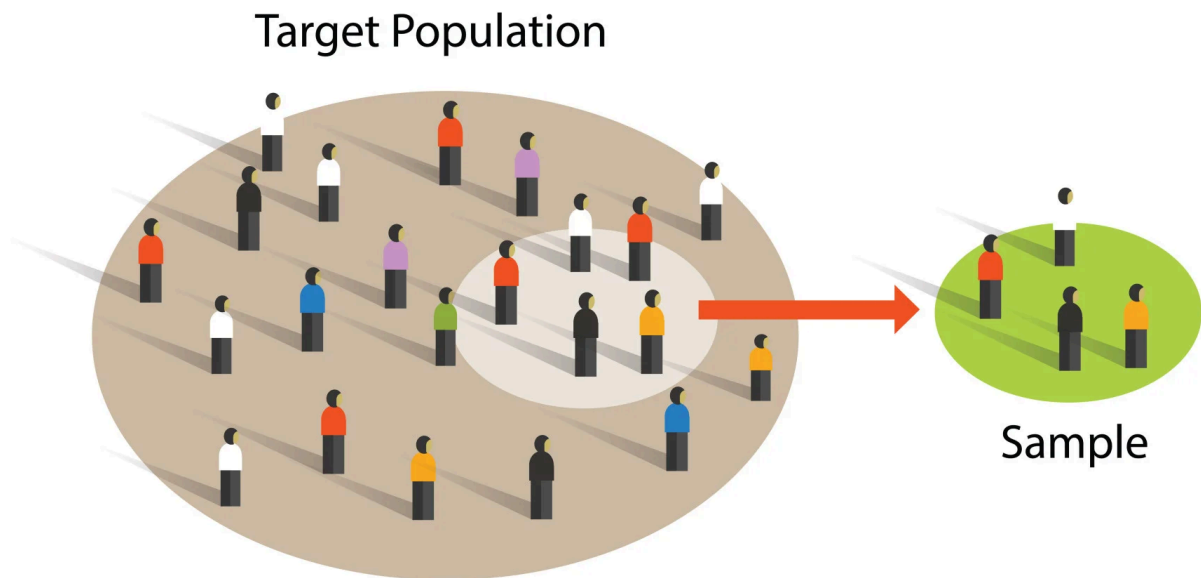
- Το τμήμα εκείνο του πληθυσμού στο οποίο έχουμε πρόσβαση (π.χ. Αν πρόκειται για άτομα, διαθέτουν κινητό τηλέφωνο) και από το οποίο μπορούμε να επιλέξουμε δεδομένα για το δείγμα



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)

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

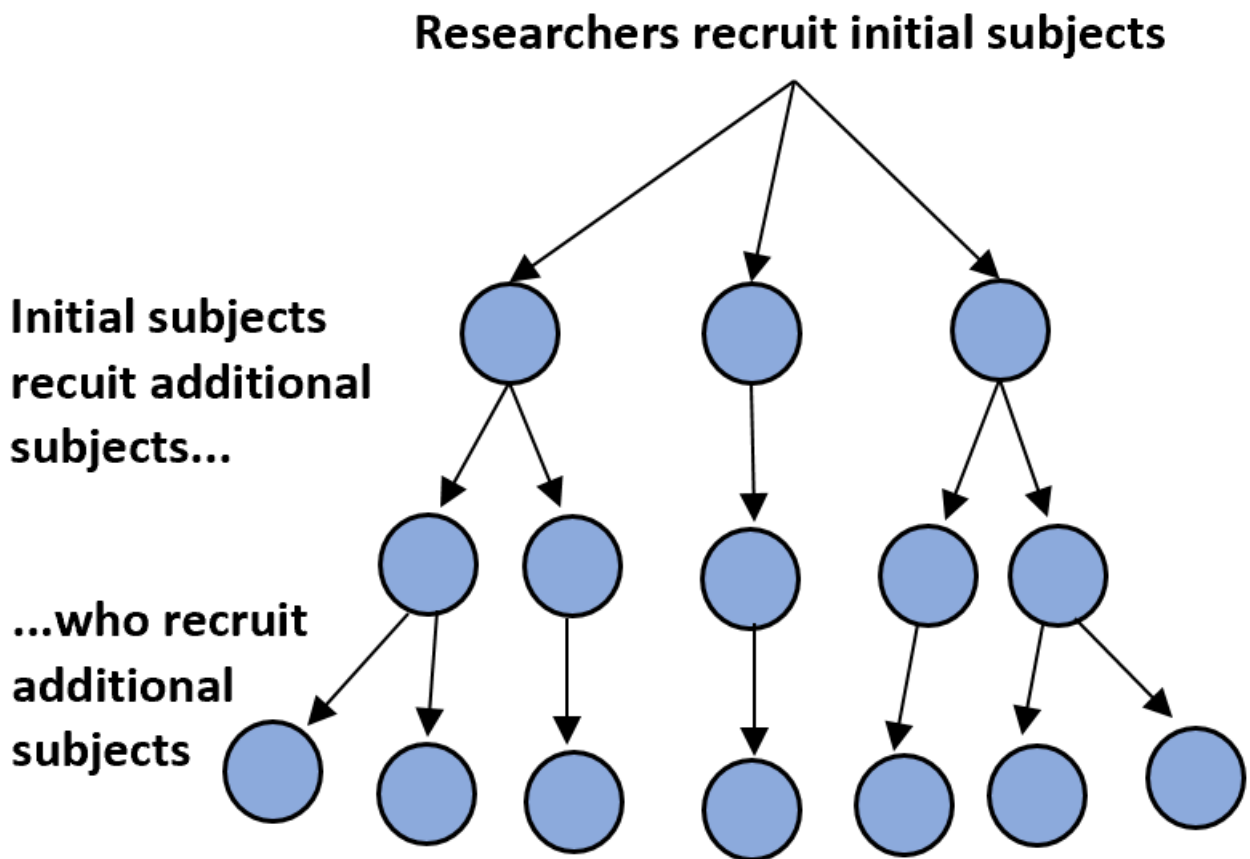


*Convenience Sampling:
Use results that are easy to get.*

Δειγματοληψία από την πρώτη σειρά της τάξης

- Θα είναι το δείγμα αντιπροσωπευτικό της τάξης;

Συγγενής είναι και η **δειγματοληψία χιονοστιβάδας** (snowball sampling)



4.2. Συστηματική δειγματοληψία

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



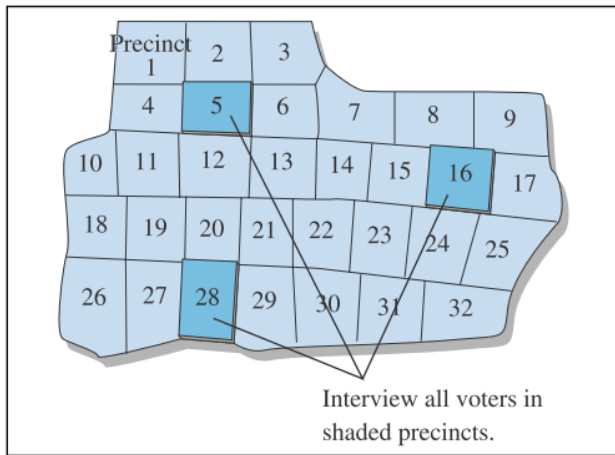
Systematic Sampling:
Select some starting point, then select every k th (such as every 50th) element in the population.

Για παράδειγμα, δειγματοληψία από όλη την τάξη

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

Ακολουθώς, επιλέγονται τυχαία μερικές από τις στιβάδες, όπως φαίνεται στην παρακάτω εικόνα



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

Για παράδειγμα, δειγματοληψία από το αριστερό και δεξιό μέρος της τάξης

4.4. Στρωματοποιημένη δειγματοληψία

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



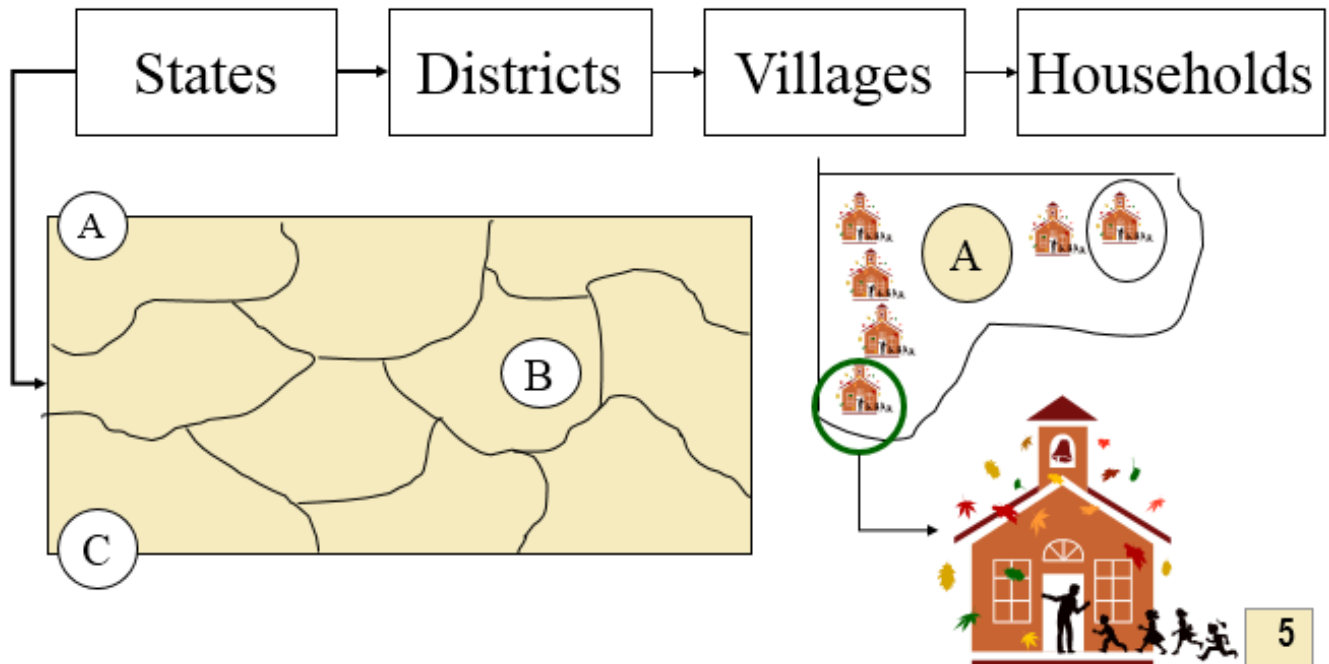
Stratified Sampling:

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.

Για παράδειγμα, δειγματοληψία βάσει ανδρών/γυναικών στην τάξη

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.

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.

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

f

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

Ας δούμε ένα **παράδειγμα** ενός μικρού πληθυσμού με πραγματικά δεδομένα των αμυντικών δαπανών (military expenditure) 10 χωρών (σε δισεκατομμύρια US\$)

Αριθμός	Χώρα	Αμυντικές δαπάνες (δισ US\$)
1	USA	801
2	China	293
3	India	110

4	Russia	87
5	UK	60
6	France	56
7	Germany	45
8	Brazil	31
9	Canada	18
10	Nigeria	5

Παίρνοντας διαφορετικά τυχαία δείγματα μεγέθους 3 από αυτόν, βλέπουμε ότι

- Μερικά μπορεί να είναι αντιπροσωπευτικά
 - China, UK, Canada
- Μερικά μπορεί να μην είναι αντιπροσωπευτικά
 - USA, China, India
 - Brazil, Canada, Nigeria

5.1. Ο νόμος των μεγάλων αριθμών

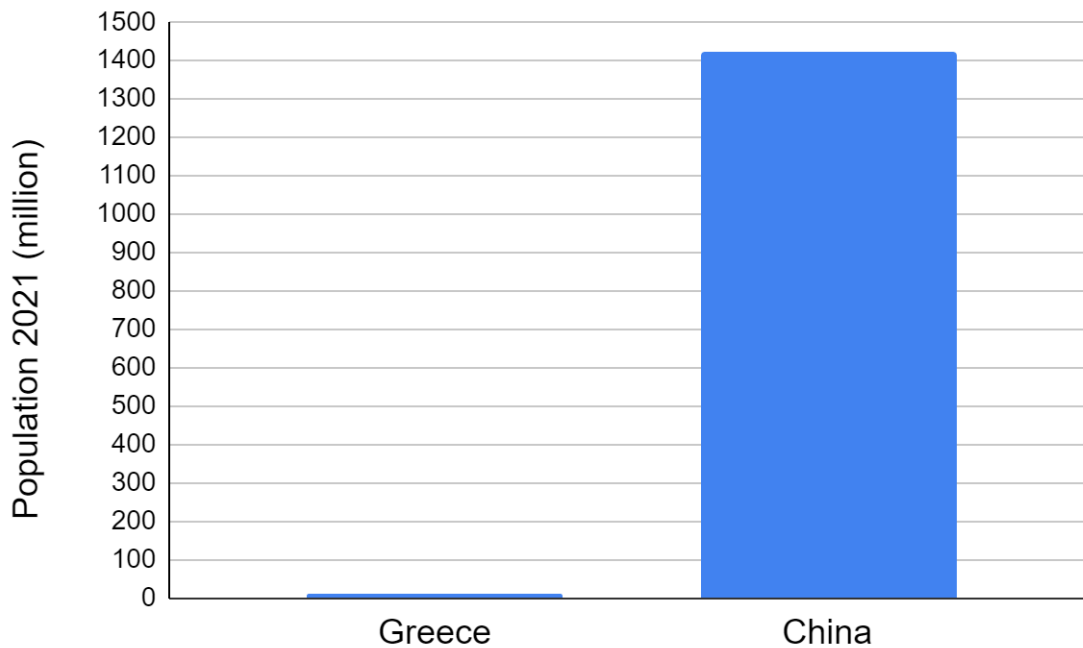
Ο **νόμος των μεγάλων αριθμών** (law of large numbers) ορίζει ότι όσο μεγαλώνει το (απόλυτο) μέγεθος ενός δείγματος (π.χ. 10 άτομα, 50 άτομα, 200 άτομα), τόσο τα δειγματικά μεγέθη (π.χ. ο μέσος όρος του δείγματος) θα προσεγγίζουν την τιμή των αντίστοιχων πληθυσμιακών παραμέτρων (π.χ. τον μέσο όρο του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

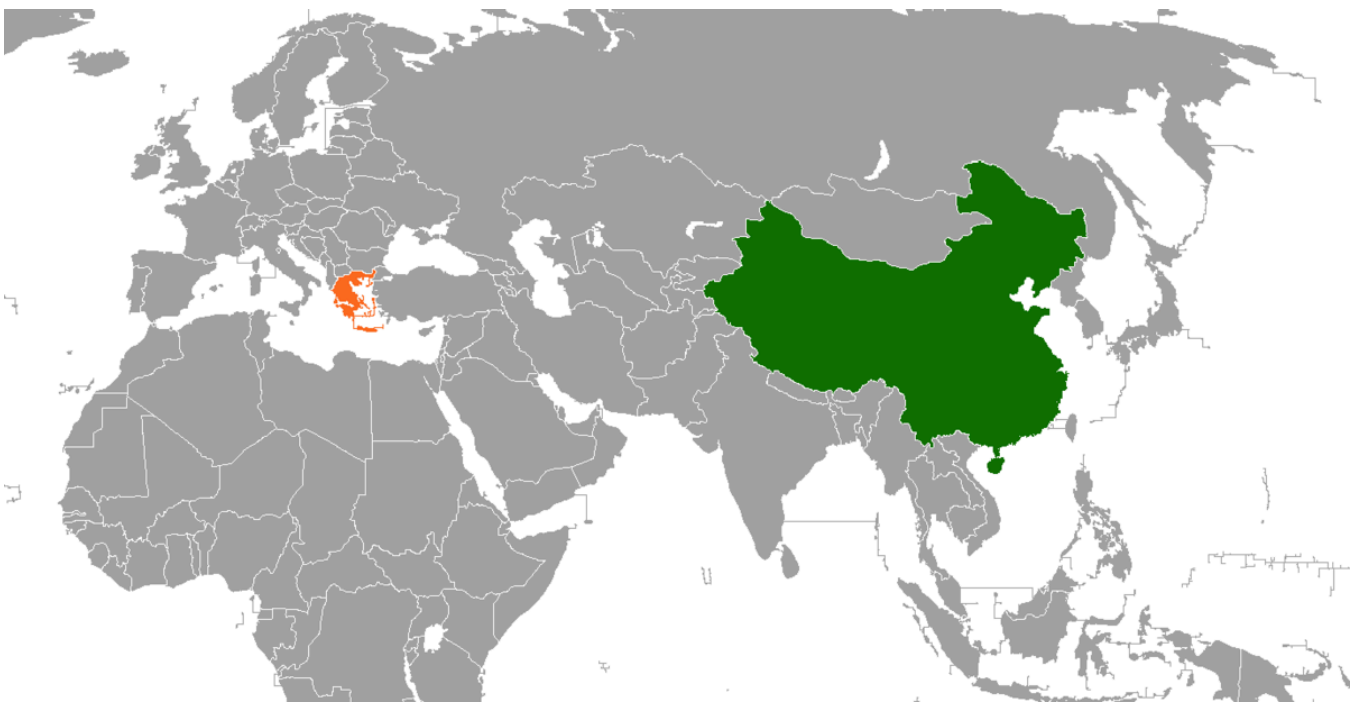
Πληθυσμός Ελλάδας και Κίνας

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

- Πληθυσμός Ελλάδας (2021) = 10.64 εκατομμύρια
- Πληθυσμός Κίνας (2021) = 1.412 δισεκατομμύρια = 1412 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 οποίο ελήφθη (10.64 εκατομμύρια για την Ελλάδα, 1421 εκατομμύρια για την Κίνα)!
- Το ότι το δείγμα των 10 χιλιάδων ατόμων αποτελεί το $10,000 / 10,640,000 \times 100 = \mathbf{0.094\%}$ (δηλαδή περίπου 0.1%) της Ελλάδας, και το $10,000 / 1,412,000,000 \times 100 = \mathbf{0.00071\%}$ 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

***** Για τις έννοιες της ενότητας αυτής, βασιστείτε στο βιβλίο σας *****

Η ενότητα αυτή είναι υπό ανακατασκευή

Total Statistical Error = Sampling Error + Nonsampling Error

Sampling Error = The "Luck of the Draw" (Cause: Natural variation)

Example: You flip a fair coin 10 times. You get 6 heads (60%) instead of the expected 5 (50%). That 10% difference is sampling error.

Nonsampling Error= The "Human Element" (Cause: Poorly worded questions, biased researchers, data entry mistakes, or people lying on surveys)

Common Types:

(a) Non-response Error: People with specific views refuse to answer.

(b) Measurement Error: A scale is calibrated incorrectly, or a question is confusing.

(c) Coverage Error: Your list of phone numbers excludes people who only use cell phones.

Increasing sample size does not fix nonsampling error

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 **δειγματικό** 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

STATISTICAL BIAS

Statistical bias refers to an error that has caused the sample to not represent the population. This error means the sample data is different from the target population.

OVERVIEW

Statistical bias refers to a systematic error or deviation in the collection, analysis, or interpretation of data that leads to consistently inaccurate or misleading results. It occurs when there is a tendency for the data to consistently overestimate or underestimate the true value being measured. It can arise from various sources, such as sampling methods and measurement techniques.

EXAMPLES

- Response Bias:** Arises when survey respondents provide inaccurate or misleading information due to factors such as social desirability, memory recall, or reluctance to disclose certain details.
- Measurement Bias:** Occurs when the measurement instrument or technique systematically misrepresents or distorts the true value being measured.

HELPFULPROFESSOR.COM

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:

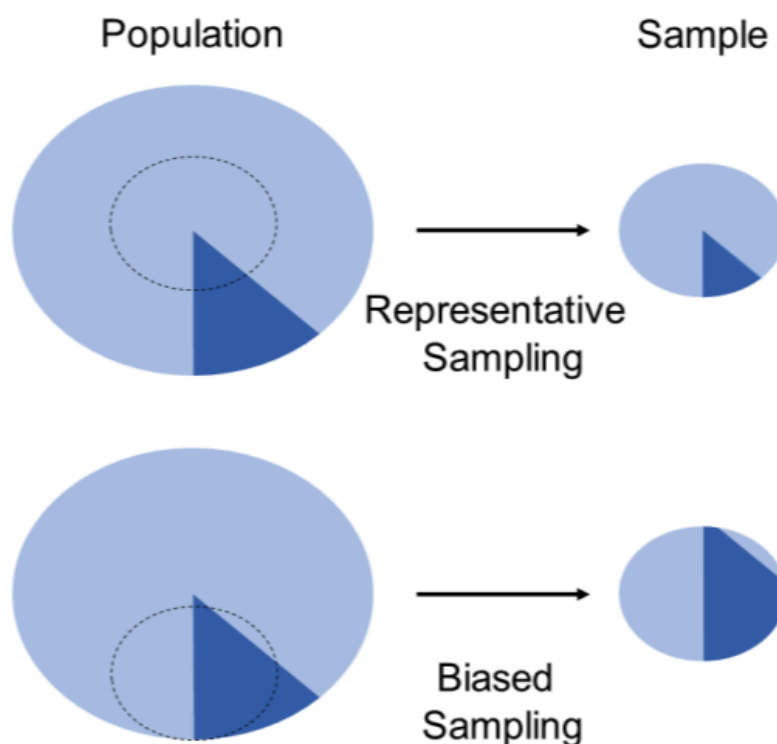
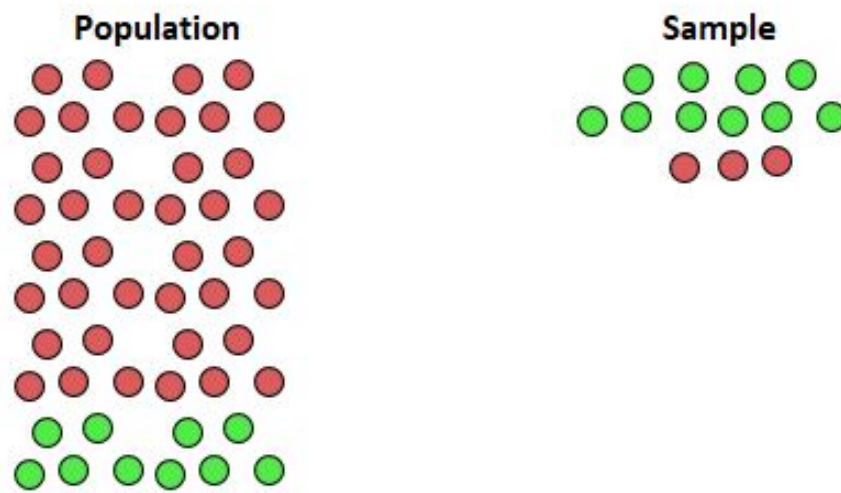
*In statistics, the term **bias** is used for describing several different concepts:*

- 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 εξαγω 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)

Standing & running pulse δείγματος φοιτητών

- Φύλο φοιτητών: τρόπος καταγραφής
- Κάπνισμα: τρόπος καταγραφής ιδιότητας καπνιστή
 - Ψευδομεταβλητές (dummy variables)

CASE	NAME	SPULSE	RPULSE	DIFF	SMOKING	MALE
1	Αφροδίτη	54	94	40	0	0
2	Βασίλης	76	100	24	1	1
3	Βικτωρία	78	110	32	0	0
4	Γιώργος	100	110	10	1	1
5	Δανάη	102	156	54	0	0
6	Δημήτρης	56	96	40	0	1
7	Ευσεβία	84	140	56	0	0
8	Κωνσταντίνος	90	92	2	0	1

9	Μαρίζα	84	108	24	0	0
---	--------	----	-----	----	---	---

6.1. Ποσοτικά δεδομένα

Είδη ποσοτικών δεδομένων

1. Διακριτά (discrete)

- Ακέραιοι αριθμοί
 - Π.χ. αριθμός παιδιών μιας οικογένειας, αριθμός φοιτητών σε ένα έτος

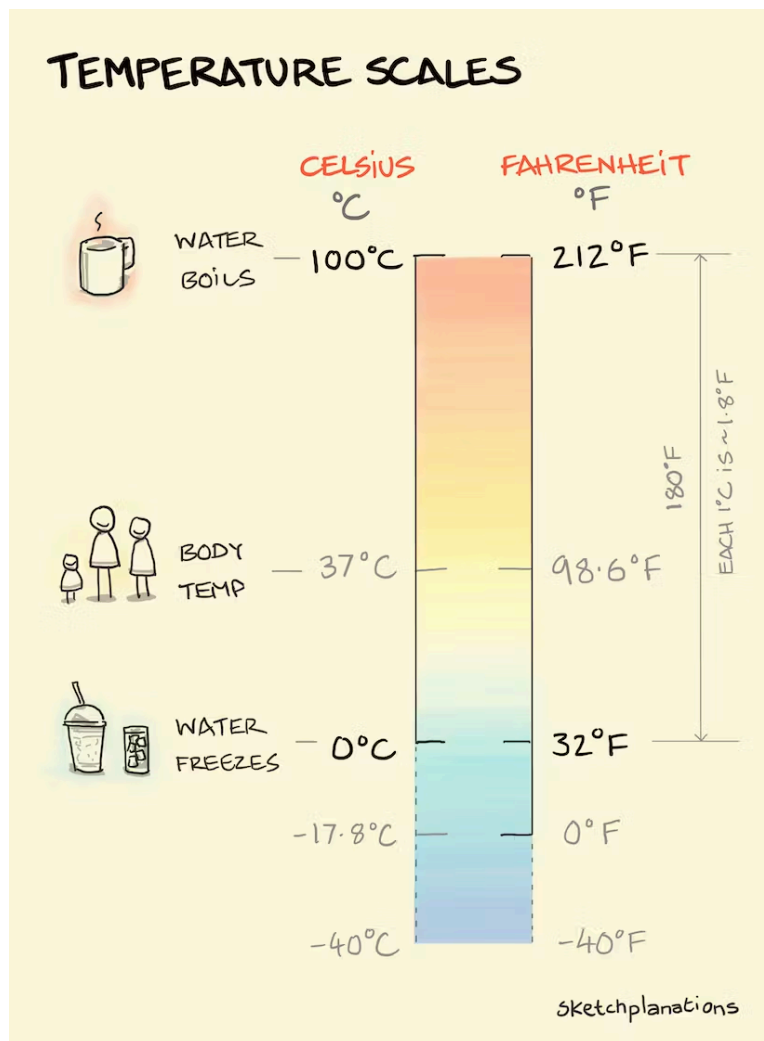
2. Συνεχή (continuous)

- 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 ($^{\circ}\text{C}$ ή $^{\circ}\text{F}$)



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 ($=104^{\circ}\text{F}$) δεν είναι δύο φορές 20°C ($=68^{\circ}\text{F}$)
 - 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. **Ταξινομικά ή 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 πανεπιστήμια (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 άνδρα

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.

6.3. Πρακτική διάκριση δεδομένων

Παρουσιάζεται εδώ πρακτική διάκριση των δεδομένων, στις κατηγορίες που χρησιμοποιούνται στην ανάλυση δεδομένων με στατιστικά προγράμματα (όπως το JASP ή το SPSS):

1. Κατηγορικά (categorical) δεδομένα ~ ποιοτικά δεδομένα

- Ονομαστικά κατηγορικά (**nominal** categorical) δεδομένα, που δεν έχουν καμία ιεράρχηση, απλά αποτελούν διαφορετικές κατηγορίες, π.χ.
 - Εθνικότητα, π.χ. Έλληνας, Κύπριος, Τούρκος, Ισραηλινός
 - Ομάδα, π.χ. Ολυμπιακός, Παναθηναϊκός, ΑΕΚ, ΠΑΟΚ
- Ταξινομικά κατηγορικά (**ordinal** categorical) δεδομένα, που έχουν μια φυσική ιεράρχηση (παρά το γεγονός ότι είναι ποιοτικά, π.χ.

- Επίκουρος καθηγητής, αναπληρωτής καθηγητής, (πρωτοβάθμιος ή τακτικός) καθηγητής
- Στρατιώτης, υποδεκανέας, δεκανέας, λοχίας, κλπ.

2. Δεδομένα κλίμακας (**scale**) ~ ποσοτικά δεδομένα, π.χ.

- Θερμοκρασία
- Εισόδημα
- Βάρος

7. Μέτρα θέσεως, διασποράς και συσχέτισης

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 = $9 - 1 = 8$

Κώδικας R για την εύρεση του ελάχιστου και μέγιστου:

```
dat <- c(1, 3, 3, 5, 6, 6, 6, 7, 9)

cat("Minimum value =", min(dat), "\n")
cat("Maximum value =", max(dat), "\n")
```

Εκτελώντας το πρόγραμμα αυτό στο R παίρνουμε

Minimum value = 1
Maximum value = 9

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

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

Τιμή (value)	Συχνότητα (frequency)	%	Πιθανότητα (probability)
1	1	11.1	0.111
3	2	22.2	0.222
5	1	11.1	0.111
6	3	33.3	0.333
7	1	11.1	0.111
9	1	11.1	0.111
Σύνολο	9	100	0.999

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% λόγω στρογγυλοποίησης)

- Η δεξιά στήλη εμφανίζει τις συχνότητες αναγόμενες στη μονάδα, δηλαδή τις πιθανότητες (probabilities), για τις οποίες θα μιλήσουμε σε επόμενη ενότητα

Κώδικας R για τον υπολογισμό των συχνοτήτων και αντίστοιχων ποσοστών (%) κάθε ακέραιας τιμής

```
dat <- c(1, 3, 3, 5, 6, 6, 6, 7, 9)

counts <- table(dat)
percents <- round(100 * prop.table(counts), 2)

print(counts)
print(percents)
```

που αν εκτελεστεί δίνει

```
dat
1 3 5 6 7 9
1 2 1 3 1 1
dat
      1      3      5      6      7      9
11.11 22.22 11.11 33.33 11.11 11.11
```

Μετά από κάθε dat, στην πρώτη γραμμή εμφανίζονται οι τιμές και στην κάτω γραμμή οι συχνότητες ή το ποσοστό κάθε τιμής

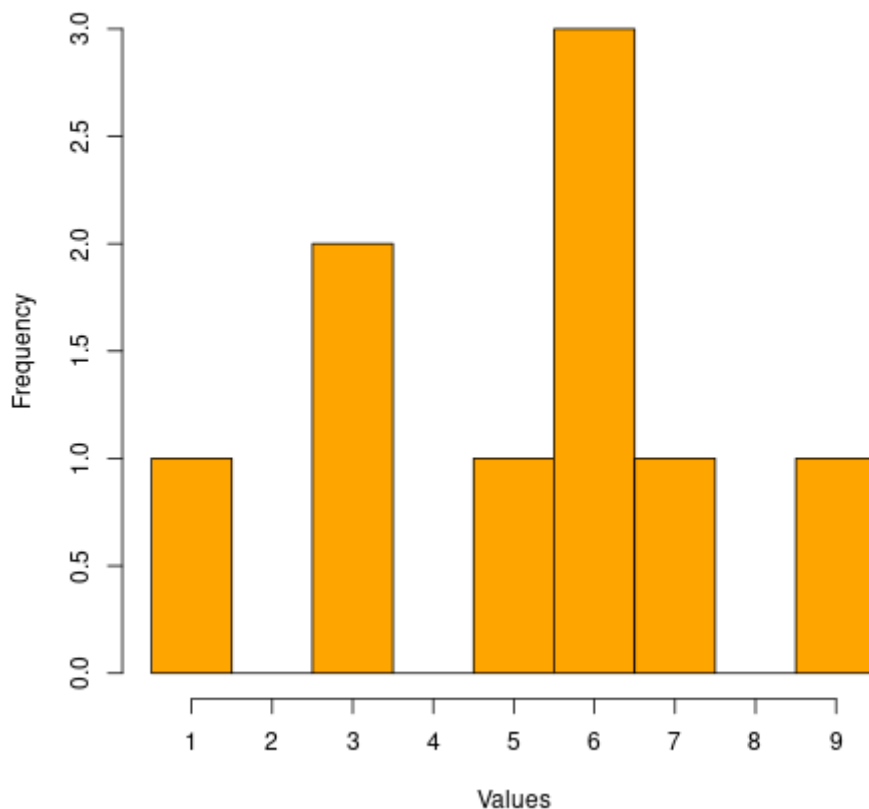
Το αντίστοιχο ιστόγραμμα στο R σχεδιάζεται με τον ακόλουθο κώδικα

```
dat <- c(1, 3, 3, 5, 6, 6, 6, 7, 9)

hist(dat,
      breaks = seq(min(dat) - 0.5, max(dat) + 0.5, by = 1),
      col = "orange",
      main = "",
      xlab = "Values",
      xaxt = "n")

axis(1, at = min(dat):max(dat))
```

και έχει ως ακολούθως



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. Διάμεσος, επικρατούσα τιμή, και τεταρτημόρια

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 $N = 6$

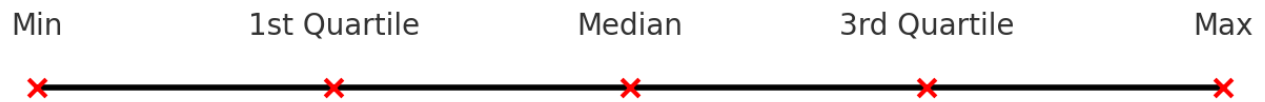
Θα υπολογίσουμε τα ακόλουθα:

- Ελάχιστο (min)

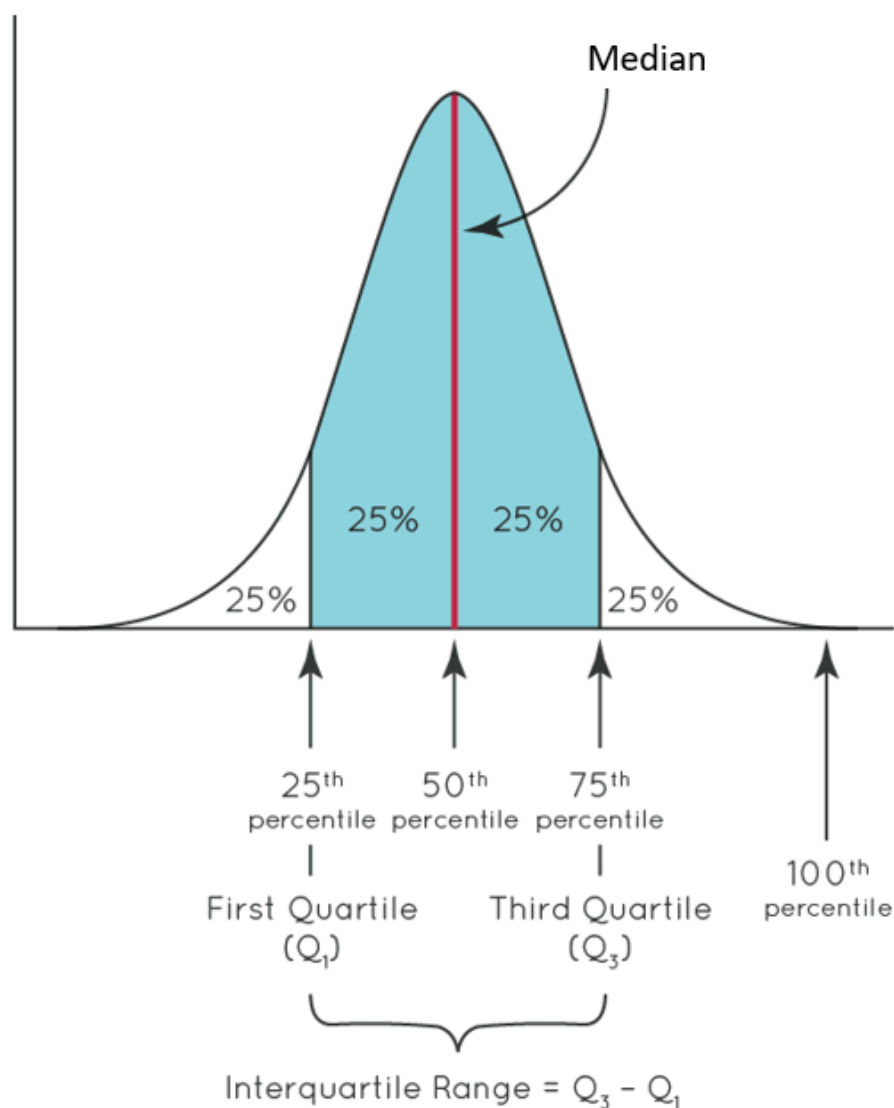
- **Διάμεση** τιμή (median ή Q_2)
 - Η τιμή που είναι στη μέση των ταξινομημένων δεδομένων, δηλαδή κάτω από την οποία βρίσκεται το 50% των δεδομένων
 - Για μεγάλα δείγματα, πιο ακριβή αποτελέσματα δίνει το να θεωρήσουμε ότι η διάμεσος είναι στη θέση $(N+1) \times 1/2$ των ταξινομημένων δεδομένων
 - Εάν το $(N+1) \times 1/2$ δεν αντιστοιχεί σε ακέραιο αριθμό, παίρνουμε το ημιάθροισμα των δύο γειτονικών δεδομένων
 - Η διάμεση τιμή είναι το δεύτερο τεταρτημόριο
- 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)
- **Πρώτο τεταρτημόριο** (first quartile ή Q_1)
 - Η τιμή κάτω από την οποία βρίσκεται το 25% των ταξινομημένων δεδομένων
 - Εάν ο αριθμός των δεδομένων μας είναι περιττός, θεωρούμε ότι το μεσαίο δεδομένο (η διάμεσος) ανήκει και στο πρώτο μισό και στο δεύτερο μισό των δεδομένων (Tukey's Hinges)
 - Για μεγάλα δείγματα, πιο ακριβή αποτελέσματα δίνει το να θεωρήσουμε ότι το Q_1 είναι στη θέση $(N+1) \times 1/4$ των ταξινομημένων δεδομένων
 - Εάν το $(N+1) \times 1/4$ δεν αντιστοιχεί σε ακέραιο αριθμό, παίρνουμε το ημιάθροισμα των δύο γειτονικών 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 ταξινομημένων δεδομένων
 - Εάν ο αριθμός των δεδομένων μας είναι περιττός, θεωρούμε ότι το μεσαίο δεδομένο (η διάμεσος) ανήκει και στο πρώτο μισό και στο δεύτερο μισό των δεδομένων (Tukey's Hinges)
 - Για μεγάλα δείγματα, πιο ακριβή αποτελέσματα δίνει το να θεωρήσουμε ότι το Q_3 είναι στη θέση $(N+1) \times 3/4$ των ταξινομημένων δεδομένων
 - Εάν το $(N+1) \times 3/4$ δεν αντιστοιχεί σε ακέραιο αριθμό, παίρνουμε το ημιάθροισμα των δύο γειτονικών δεδομένων

- Το τρίτο τεταρτημόριο είναι η διάμεσος του δεύτερου μισού!
- **Ενδοτεταρτημοριακό εύρος** (interquartile range ή IQR)
 - Ισούται με $Q_3 - Q_1$

Relative Position of Five-Number Summary



Quartiles and Percentiles



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	<i>min</i>
76.5	2	76.5	Q_1
76.8	3	76.7	<i>median</i>
76.8	4	76.8	
87.6	5	76.8	Q_3
76.7	6	87.6	<i>max</i>

Δουλεύοντας στα ταξινομημένα δεδομένα, υπολογίζουμε τα παρακάτω:

- Ελάχιστο = 75.5
- Μέγιστο = 87.6
- Εύρος = Μέγιστο - Ελάχιστο = $87.6 - 75.5 = 12.1$
- Διάμεσος
 - Ως το ημιάθροισμα των δύο δεδομένων που είναι στη μέση:
 $(76.7 + 76.8) / 2 = 76.75$
 - 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 $IQR = Q_3 - Q_1 = 76.8 - 76.5 = 0.3$

Descriptive Statistics

	LENGTH
Valid	6
Missing	0
Mode	76.76
Median	76.75
Mean (arithmetic)	78.32
Std. Deviation	4.574
Skewness	2.383
Std. Error of Skewness	0.845
Range	12.10
Minimum	75.50
Maximum	87.60
25th percentile	76.55
50th percentile	76.75
75th percentile	76.80

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

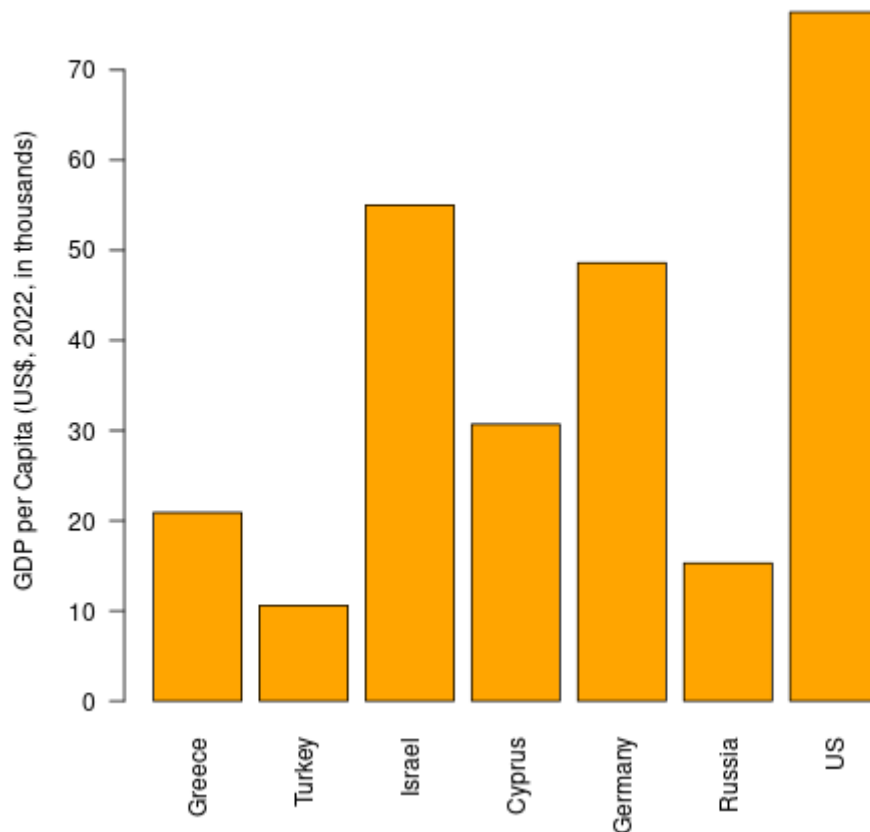
Παρατήρηση	Country	GDP pc (US\$ 2022, 1000s)
1	Greece	20.9
2	Turkey	10.6
3	Israel	55
4	Cyprus	30.7
5	Germany	48.6
6	Russia	15.3

7	US	76.4
---	----	------

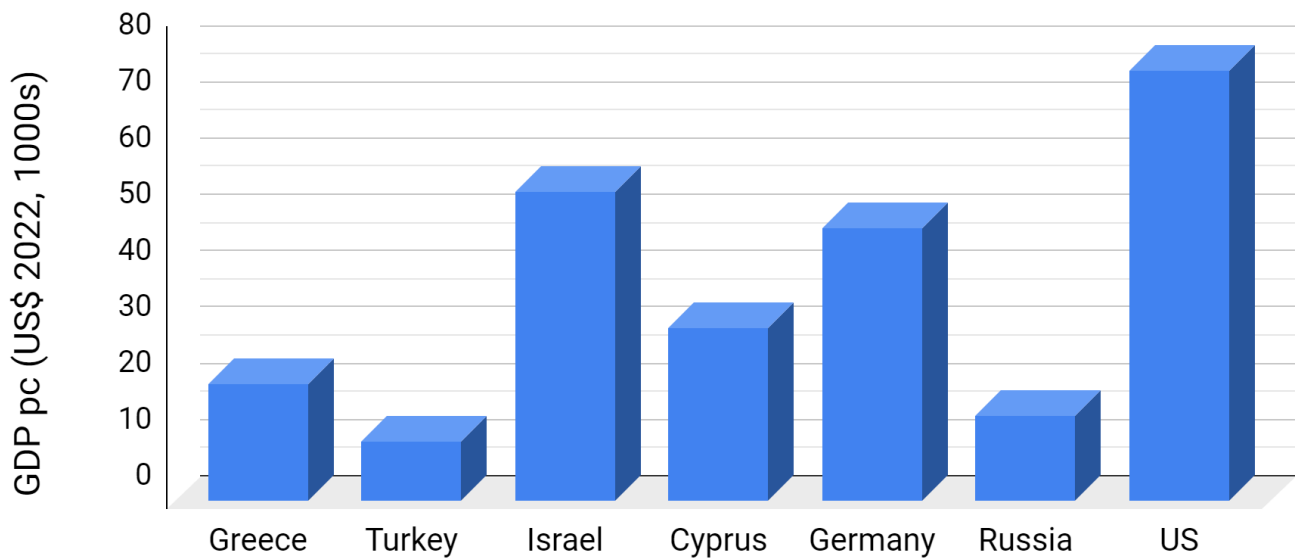
Τα δεδομένα αυτά μπορούμε να τα απεικονίσουμε σε ραβδόγραμμα (bar chart) με τον παρακάτω κώδικα R

```
countries <- c("Greece", "Turkey", "Israel", "Cyprus",
               "Germany", "Russia", "US")
gdp_pc <- c(20.9, 10.6, 55.0, 30.7, 48.6, 15.3, 76.4)

barplot(gdp_pc,
        names.arg = countries,
        ylab = "GDP per Capita (US$, 2022, in thousands)",
        col = "orange",
        las = 2) # Rotates country labels for readability
```



Τρισδιάστατα διαγράμματα, όπως το παρακάτω, **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

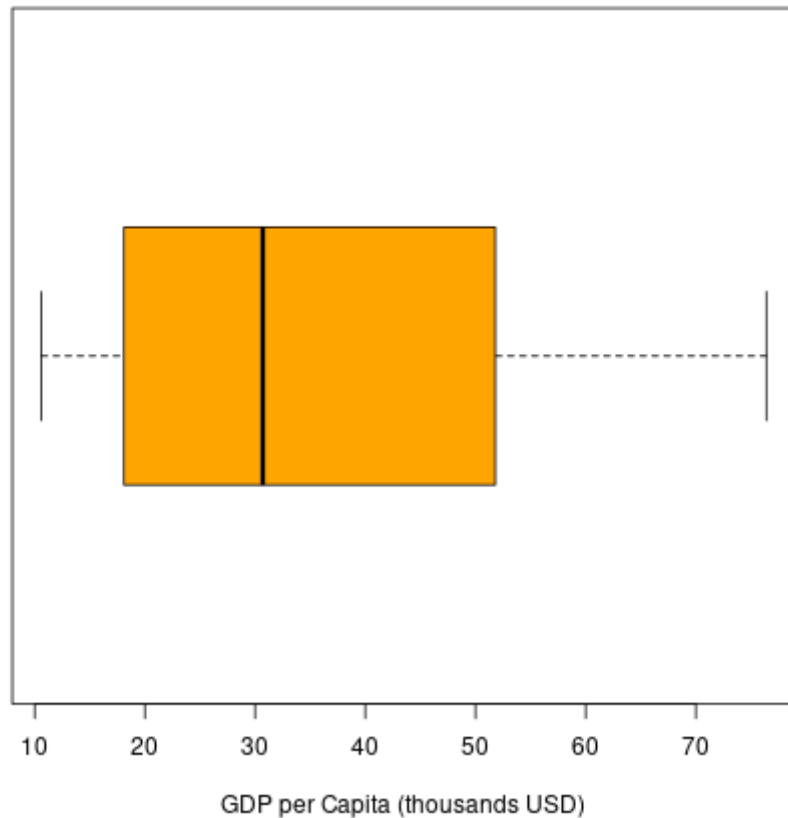
Υπολογισμοί

- 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 $Q_1 = (15.3 + 20.9) / 2 = 18.1$
 - Επειδή έχουμε περιττό αριθμό δεδομένων, θεωρούμε ότι η διάμεσος ανήκει (και) στο πρώτο μισό (και στο δεύτερο μισό)
- Τρίτο τεταρτημόριο $Q_3 = (48.6 + 55) / 2 = 51.8$
 - Επειδή έχουμε περιττό αριθμό δεδομένων, πάλι θεωρούμε ότι η διάμεσος ανήκει (και στο πρώτο μισό και) στο δεύτερο μισό

- Ενδοτεταρτημοριακό εύρος $IQR = Q_3 - Q_1 = 51.8 - 18.1 = 33.7$

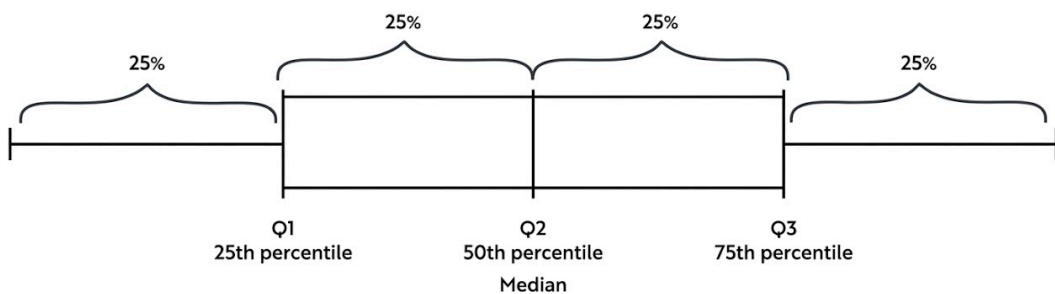
Descriptive Statistics

	GPDpc
Valid	7
Missing	0
Mode	21.76
Median	30.70
Mean (arithmetic)	36.79
Std. Deviation	24.07
IQR	33.70
Skewness	0.631
Std. Error of Skewness	0.794
Range	65.80
Minimum	10.60
Maximum	76.40
25th percentile	18.10
50th percentile	30.70
75th percentile	51.80



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

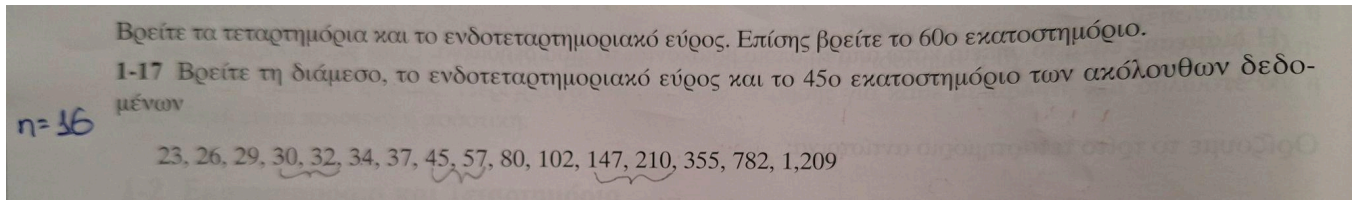


Γι' αυτό το boxplot ονομάζεται και box and whiskers plot

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

Από σύγγραμμα μαθήματος



Και ένα χιουμοριστικό παράδειγμα

- Σε ποιο μέτρο θέσεως αναφέρεται ο George Carlin σε αυτό το βίντεο:
<https://youtu.be/AKN1Q5Sjbel?t=22>

7.4. Εσωτερικοί και εξωτερικοί φράχτες

Εμπειρικός 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

Εσωτερικοί και εξωτερικοί φράχτες

Πηγή παραδείγματος [εδώ](#)

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

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
35	900
36	982
37	983
38	985
39	999
40	1001
41	1002
42	1003
43	1009
44	1100

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
45	1101
46	1102
47	1103
48	1109
49	1201
50	1301
51	1399
52	1400
53	1501
54	1599

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
55	1699

Πλήθος δεδομένων: $n=55$

Δεδομένου ότι το δείγμα είναι μεγάλο, αυτή τη φορά ας βρούμε τη θέση της διαμέσου και των τεταρτημορίων με τον τρόπο που αναφέραμε προηγουμένως

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

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

- Το 28 είναι η θέση της διαμέσου, όχι η τιμή της!
- Άρα, η διάμεσος (median) θα είναι το δεδομένο που βρίσκεται στη θέση 28, δηλαδή το 703

- Εάν ο τύπος για τη θέση της διαμέσου μας δώσει δεκαδικό αποτέλεσμα, π.χ. 10.2, θα πρέπει να πάρουμε το ημιάθροισμα των σημείων που βρίσκονται πριν και μετά από την τιμή αυτή, π.χ. το 10ο και 11ο σημείο

(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 : $(n+1) \times 1/4 = (n+1)/4 = (55+1)/4 = 56/4 = 14$

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

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

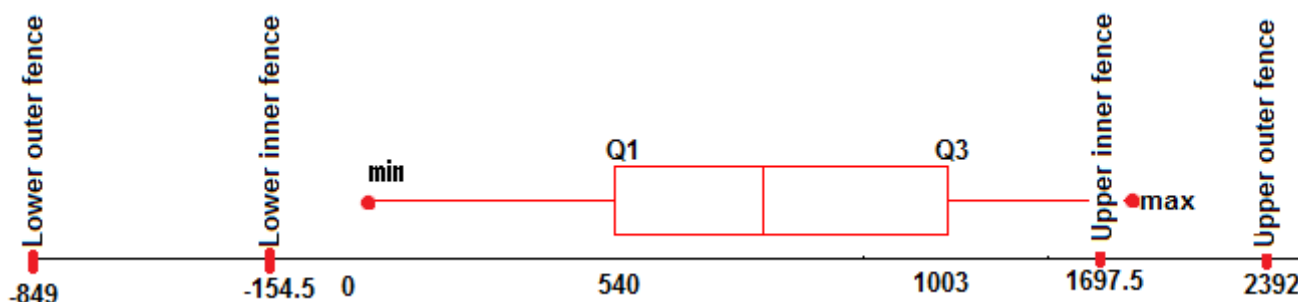
- Επομένως, το Q_3 θα είναι το δεδομένο που βρίσκεται στη θέση 42, δηλαδή το 1003

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

Υπολογισμός φραχτών

- Lower outer fence: $Q_1 - 3 \times IQR = 540 - 3 \times 463 = -849$
- Lower inner fence: $Q_1 - 1.5 \times IQR = 540 - 1.5 \times 463 = -154.5$
- (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: $Q_3 + 1.5 \times IQR = 1003 + 1.5 \times 463 = 1697.5$
- Upper outer fence: $Q_3 + 3 \times IQR = 1003 + 3 \times 463 = 2392$

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)

Descriptive Statistics

STAT NUMBERS	
Valid	55
Missing	1
Mode	587.1
Median	703.0
Mean (arithmetic)	761.1
Std. Deviation	379.4
IQR	461.0
Skewness	0.463
Std. Error of Skewness	0.322
Range	1670
Minimum	29.00
Maximum	1699
25th percentile	541.5
50th percentile	703.0
75th percentile	1003

Οι τιμές των Q_1 and Q_3 είναι ελαφρά διαφορετικές γιατί το R χρησιμοποιεί πιο ακριβείς μεθόδους για τον υπολογισμό τους

Και για τον υπολογισμό των εσωτερικών και εξωτερικών φρακτών, ο κώδικας R

```
data <- c(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)

qs <- quantile(data, probs = c(0.25, 0.75))
```

```
iqr <- qs[2] - qs[1]

lower_inner <- qs[1] - 1.5 * iqr
upper_inner <- qs[2] + 1.5 * iqr

lower_outer <- qs[1] - 3 * iqr
upper_outer <- qs[2] + 3 * iqr

cat("Lower outer fence =", lower_outer, "\n")
cat("Lower inner fence =", lower_inner, "\n")
cat("Upper inner fence =", upper_inner, "\n")
cat("Upper outer fence =", upper_outer, "\n")
```

δίνει

```
Lower outer fence = -841.5
Lower inner fence = -150
Upper inner fence = 1694
Upper outer fence = 2385.5
```

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

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

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

Εκπομπές διοξειδίου του άνθρακα

από [εδώ](#): εκπομπές διοξειδίου του άνθρακα (CO₂) 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 CO₂ του έτους 2016 για αυτές τις 5 χώρες είναι 4.176 δισεκατομμύρια τόνοι

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

Για να υπολογίσουμε τη μέση τιμή των παραπάνω δεδομένων στο R γράφουμε τον παρακάτω κώδικα

```
values <- c(10.433, 5.012, 2.534, 1.662, 1.24)
cat("Mean =", mean(values), "\n")
```

Εκτελώντας τον παραπάνω κώδικα, παίρνουμε τα ακόλουθα αποτελέσματα

```
Mean = 4.1762
```

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

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

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 τυπικής απόκλισης για τις ετήσιες εκπομπές CO₂ των 5 παραπάνω χωρών

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{(10.433-4.176)^2 + (5.012-4.176)^2 + (2.534-4.176)^2 + (1.662-4.176)^2 + (1.24-4.176)^2}{5-1}} \\ &= \sqrt{\frac{(6.257)^2 + (0.836)^2 + (-1.642)^2 + (-2.514)^2 + (-2.936)^2}{4}} \\ &= \sqrt{\frac{39.15 + 0.699 + 2.696 + 6.32 + 8.62}{4}} \end{aligned}$$

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

Η τυπική απόκλιση λοιπόν των ετήσιων εκπομπών CO₂ 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

Κώδικας R για τον υπολογισμό της τυπικής απόκλισης

```
values <- c(10.433, 5.012, 2.534, 1.662, 1.24)
cat("Standard deviation =", sd(values), "\n")
```

Εκτελώντας τον παίρνουμε

Standard deviation = 3.790956

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

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

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

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

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 CO₂ 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)

Κώδικας R για τον υπολογισμό της διακύμανσης

```
values <- c(10.433, 5.012, 2.534, 1.662, 1.24)
cat("Variance =", var(values), "\n")
```

που μας δίνει

Variance = 14.37135

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

Standing pulse δείγματος φοιτητών

CASE	NAME	SPULSE	SMOKING	MALE
1	Αφροδίτη	54	0	0
2	Βασίλης	76	1	1
3	Βικτωρία	78	0	0
4	Γιώργος	100	1	1
5	Δανάη	102	0	0
6	Δημήτρης	56	0	1
7	Ευσεβία	84	0	0
8	Κωνσταντίνος	90	0	1
9	Μαρίζα	84	0	0

Κώδικας R (δείτε στο **κίτρινο highlight** πως υπολογίζουμε το CV)

```
SPULSE <- c(54, 76, 78, 100, 102, 56, 84, 90, 84)
cat("Mean =", mean(SPULSE), "\n")
cat("Standard deviation =", sd(SPULSE), "\n")
cat("Coefficient of variation (%) =", sd(SPULSE) /
mean(SPULSE) * 100, "\n")
```

που δίνει

Mean = 80.44444

Standard deviation = 16.90496
Coefficient of variation (%) = 21.01446

Ένα άλλο μέτρο διασποράς των δεδομένων είναι ο **συντελεστής μεταβλητότητας ή διασποράς** (coefficient of variation), που εκφράζεται σε ποσοστό τοις εκατό

The coefficient of variation (or CV) for a set of sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

→ Free of specific units of measure

→ For comparing variation for values taken from different populations

Sample

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

Population

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

(Η χρήση του πεζού γράμματος n για το μέγεθος του δείγματος και του κεφαλαίου γράμματος N για το μέγεθος του πληθυσμού δεν τηρείται πάντα)

Για το προηγούμενο παράδειγμα των εκπομπών CO₂

$$CV = \frac{s}{\bar{x}} \times 100 = \frac{3.791}{4.176} \times 100 = 90.78\%$$

Νέος κώδικας R που υπολογίζει και τυπώνει όλα τα μέχρι τώρα μέτρα θέσεως και διασποράς για τα δεδομένα των εκπομπών CO₂ και μάλιστα με συγκεκριμένα δεκαδικά ψηφία

```
C02 <- c(10.433, 5.012, 2.534, 1.662, 1.24)
cat("Mean =", mean(C02), "\n")
cat("Standard deviation =", sd(C02), "\n")
cat("Coefficient of variation (%) =", sd(C02) / mean(C02) *
100, "\n")
```

που δίνει

Mean = 4.1762
Standard deviation = 3.790956
Coefficient of variation (%) = 90.77526

Διασπορά δεδομένων

Data Set I	41	44	45	47	47	48	51	53	58	66
Data Set II	20	37	48	48	49	50	53	61	64	70

Υπολογισμός μέσης τιμής και τυπικής απόκλισης

Data Set I

$$n = 10$$

$$\bar{x} = 50$$

$$s = 7.41$$

$$CV = 14.82\%$$

Data Set II

$$n = 10$$

$$\bar{x} = 50$$

$$s = 14.16$$

$$CV = 28.32\%$$

Πόσα δεδομένα πέφτουν μέσα στα διαστήματα $\bar{x} \pm 3s$;

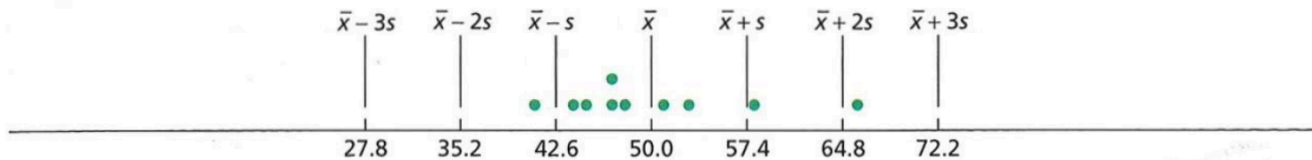
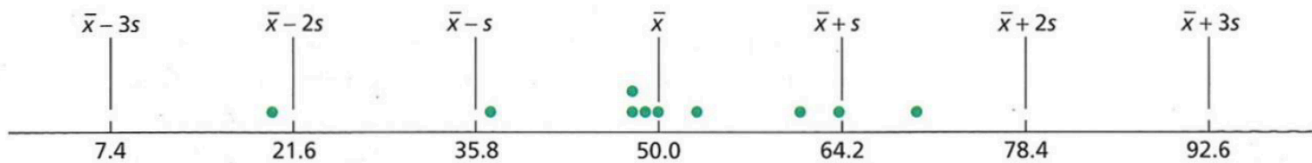
Data Set I

$$\bar{x} \pm 3s = 50 \pm 3 \times 7.41 = (27.77, 72.23)$$

Data Set II

$$\bar{x} \pm 3s = 50 \pm 3 \times 14.16 = (7.53, 92.47)$$

(Στο παρακάτω γράφημα, οι τιμές του οριζόντιου άξονα έχουν υπολογιστεί μετά από στρογγυλοποίηση των τιμών της τυπικής απόκλισης στις εικονιζόμενες)

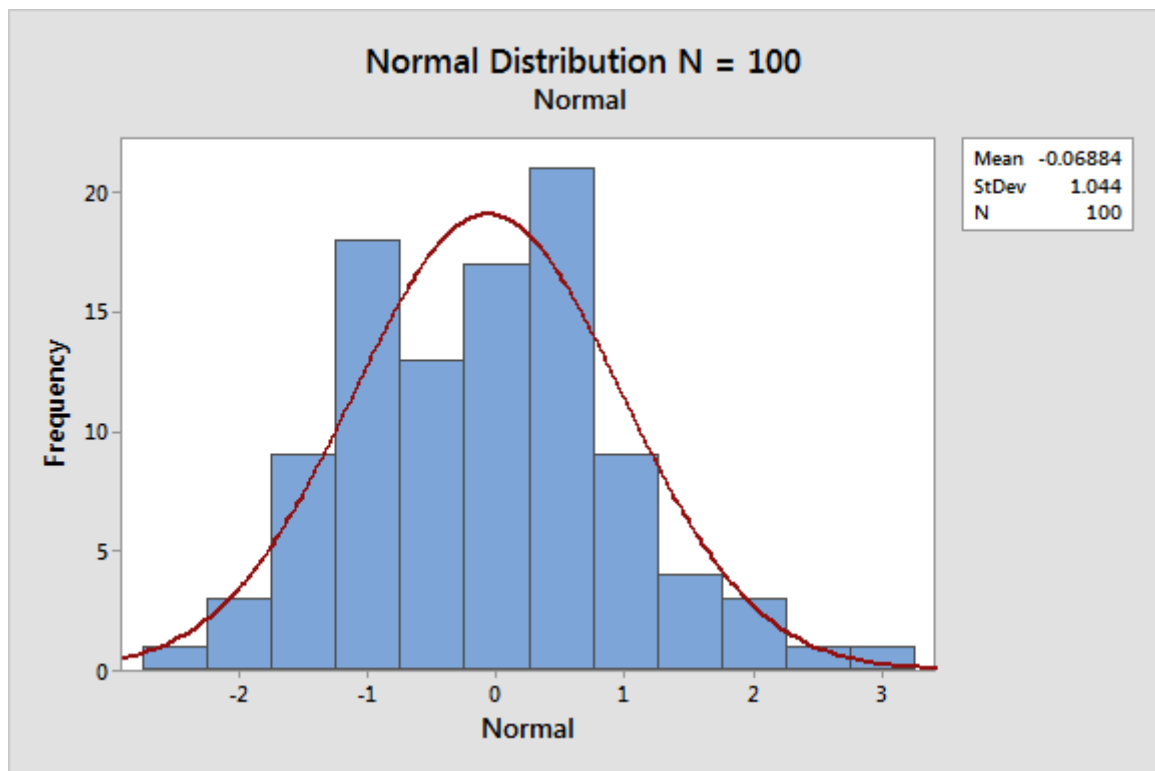
FIGURE 3.6Data Set I; $\bar{x} = 50, s = 7.4$ **FIGURE 3.7**Data Set II; $\bar{x} = 50, s = 14.2$ 

Σημαντικό συμπέρασμα

- Στα περισσότερα σύνολα δεδομένων (δείγματα), τα περισσότερα δεδομένα θα βρίσκονται μέσα στο διάστημα $\bar{x} \pm 3s$

7.6. Συμμετρία και στρέβλωση

Μια εικόνα της κατανομής ενός συνόλου δεδομένων δίνεται από το ιστόγραμμα

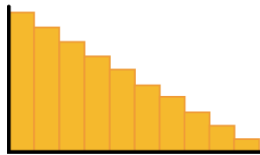


Στα παρακάτω διαγράμματα λάβετε υπόψη ότι υπάρχει ορθογραφικό λάθος (“*distriutions*” αντί για “*distributions*”)

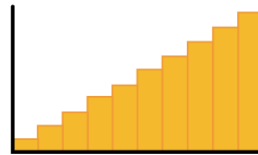
Symmetric (normal) vs skewed and uniform distributions



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



**Right-skewed
distribution**
(Positively-skewed)

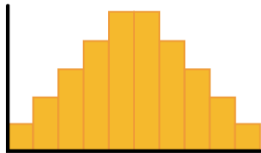


**Left-skewed
distribution**
(Negatively-skewed)

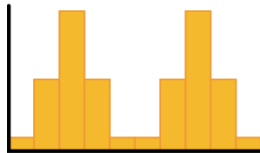


Uniform distribution
(equal spread,
no peaks)

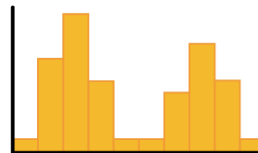
Unimodal vs bimodal distributions



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



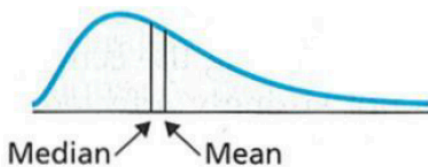
**Symmetric bimodal
distribution**
(two modes)



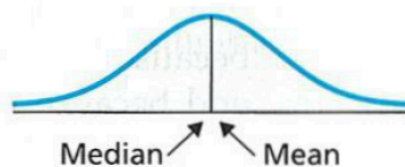
**Non-symmetric
bimodal distribution**
(two modes)

Στρέβλωση (skewness)

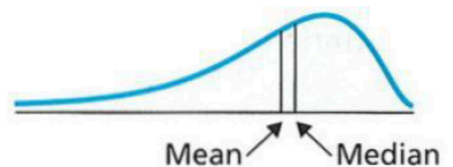
- Συμμετρικές κατανομές: μηδενική στρέβλωση
- Όταν το "σουβλί" (η λεπτή "ουρά" της κατανομής) είναι στα αριστερά: αρνητική στρέβλωση
- Όταν το σουβλί είναι στα δεξιά: θετική στρέβλωση



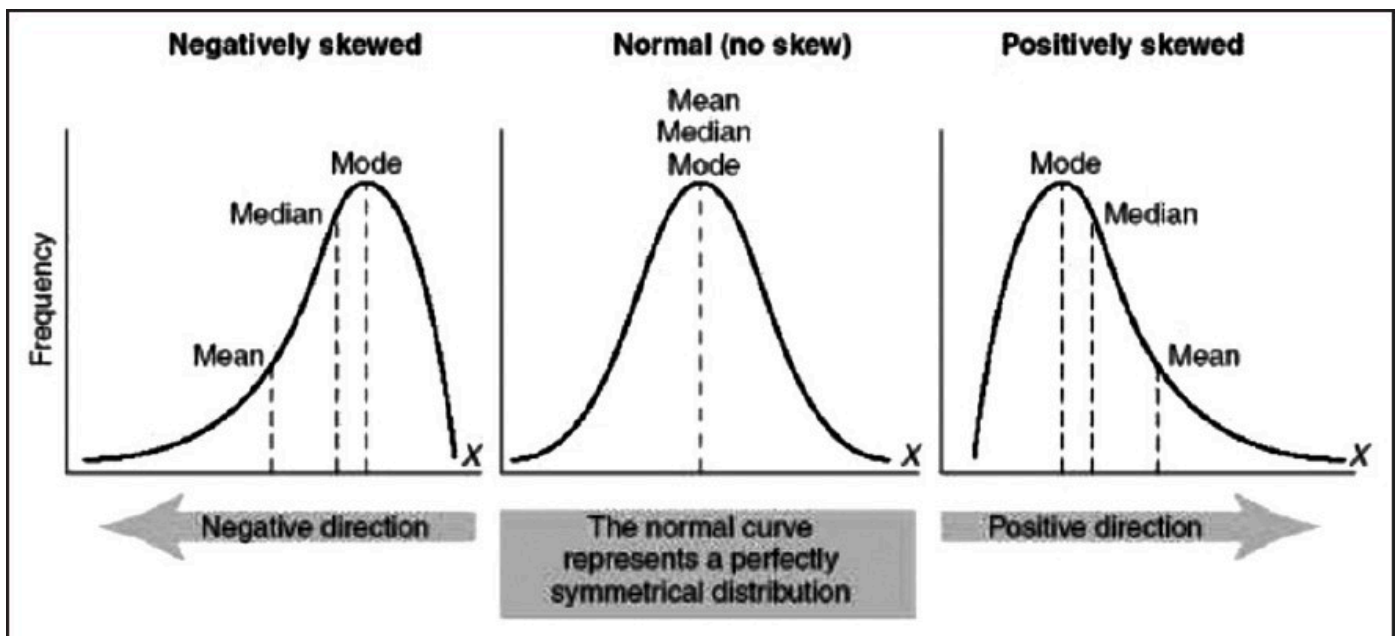
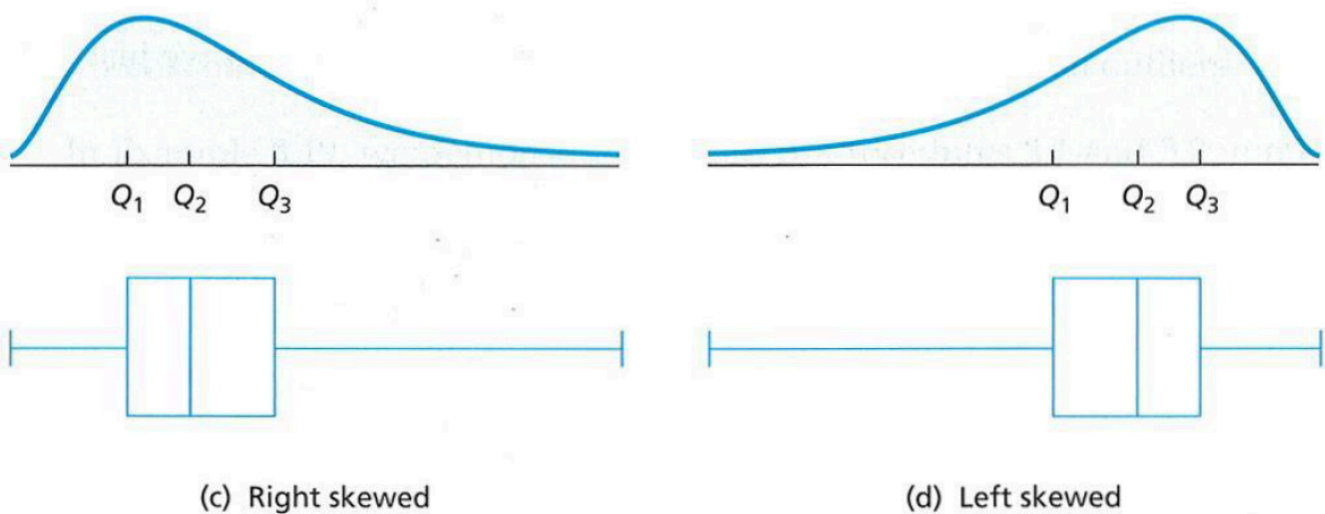
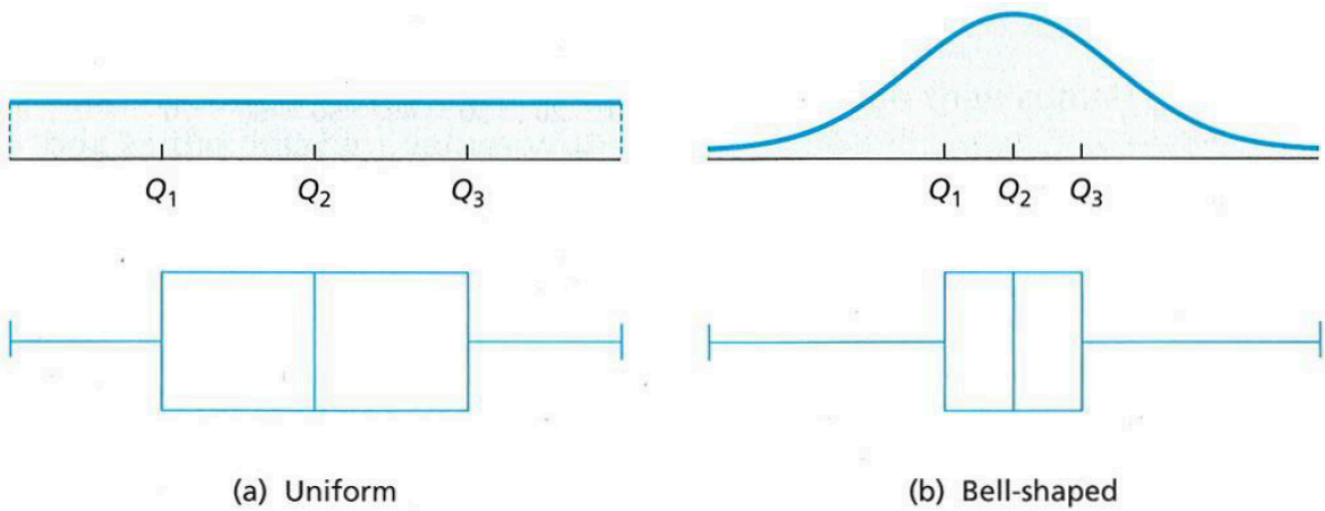
(a) Right skewed



(b) Symmetric



(c) Left skewed



7.7. Συσχέτιση

Σε αυτή την ενότητα εξετάζουμε για πρώτη φορά περισσότερες περισσότερες της μιας μεταβλητές

Δύο μεταβλητές, X και Y, μπορεί να έχουν θετική, αρνητική ή μηδενική συσχέτιση

- **Θετική** συσχέτιση σημαίνει ότι όταν αυξάνει η τιμή της μιας, εν γένει αυξάνει και η τιμή της άλλης
- **Αρνητική** συσχέτιση σημαίνει ότι όταν αυξάνει η τιμή της μιας, εν γένει μειώνεται η τιμή της άλλης
- **Μηδενική** συσχέτιση σημαίνει ότι οι μεταβολές στην τιμή της μιας δεν έχει σχέση με μεταβολές στην τιμή της άλλης

Την ισχύ της συσχέτισης τη μετράει ο **συντελεστής συσχέτισης** (correlation coefficient)

- Συμβολίζεται με το Αγγλικό γράμμα r
- $-1 \leq r \leq 1$

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

n : sample size

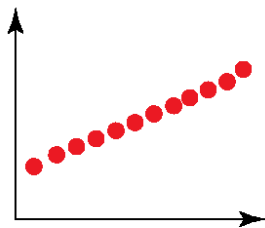
x_i, y_i : individual sample points indexed with i

$\bar{x}, \bar{y}$: sample mean

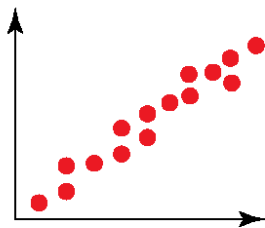
Η παραπάνω εξίσωση δεν μας ενδιαφέρει ιδιαίτερα, γιατί τον συντελεστή συσχέτισης δεν πρόκειται ποτέ να τον υπολογίσουμε με το χέρι

- Πρέπει όμως να ξέρουμε να αξιολογούμε
 - Τη φύση της συσχέτισης ανάμεσα στις δύο μεταβλητές
 - π.χ. γραμμική, μη γραμμική
 - Το πρόσημο και την τιμή του συντελεστή συσχέτισης, που περιγράφει τη συσχέτιση αυτή

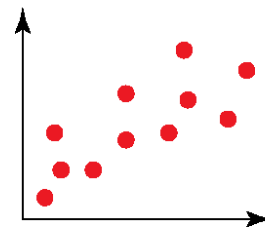
Τα παρακάτω διαγράμματα ονομάζονται **διαγράμματα διασποράς** ή **στικτογράμματα** (scatterplots)



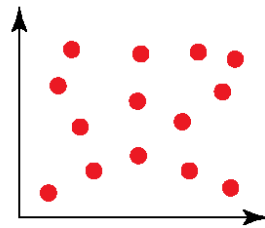
Perfect
Positive
Correlation



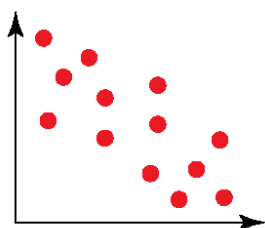
Strong
Positive
Correlation



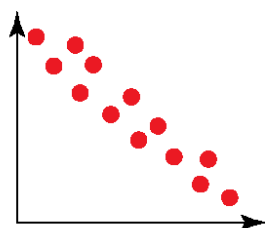
Weak
Positive
Correlation



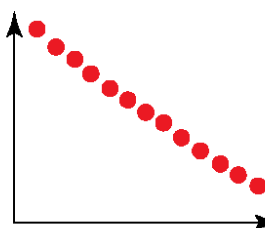
No
Correlation



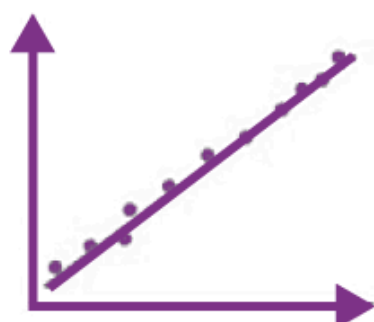
Weak
Negative
Correlation



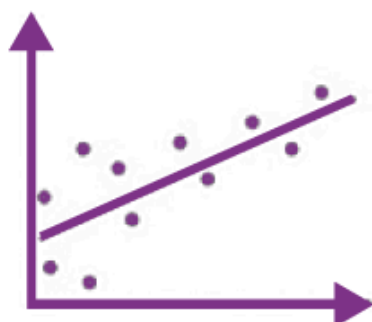
Strong
Negative
Correlation



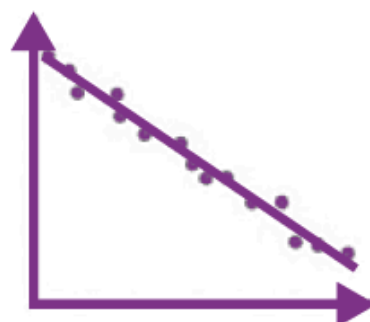
Perfect
Negative
Correlation



Strong positive correlation



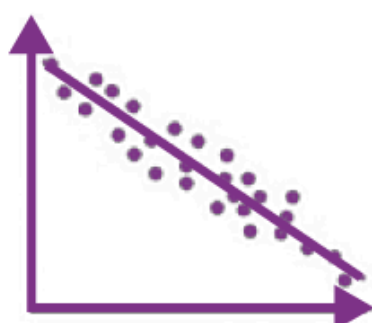
Weak positive correlation



Strong negative correlation



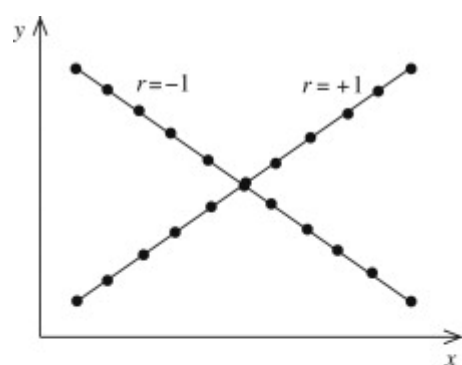
Weak negative correlation



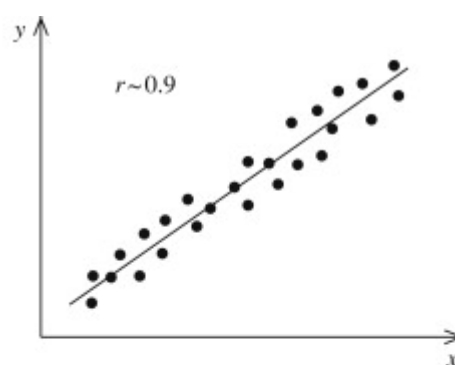
Moderate negative correlation



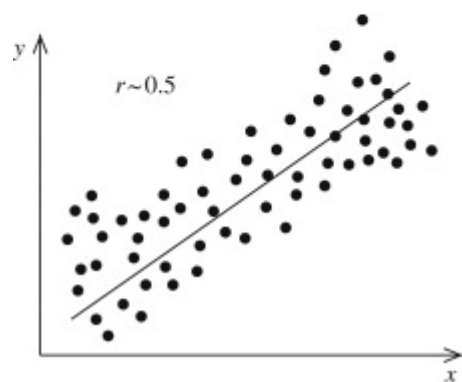
No correlation



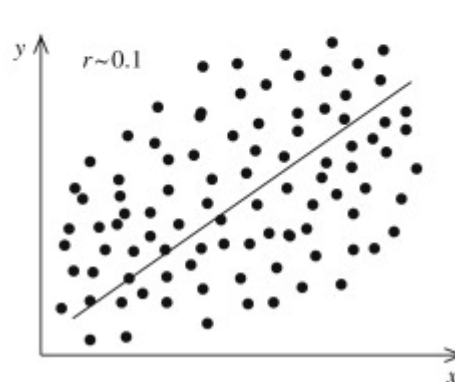
(a)



(b)

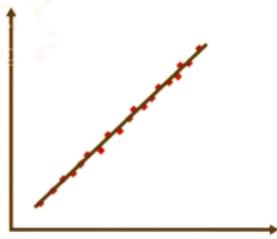


(c)

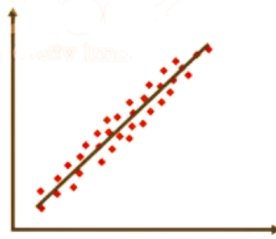


(d)

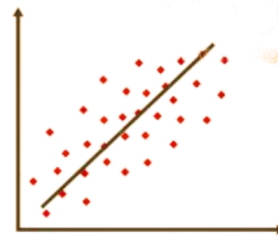
Correlation Coefficients



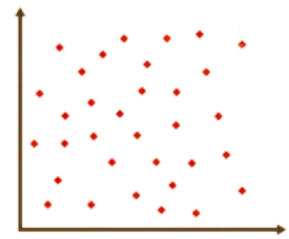
Perfect Positive
Correlation ($r = +1$)



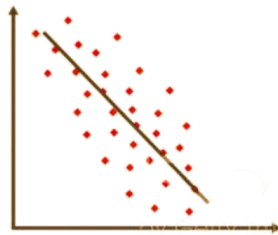
Strong Positive
Correlation ($0.7 \leq r < +1$)



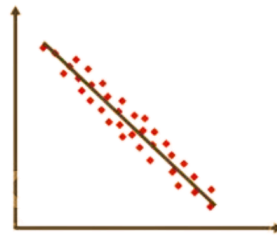
Weak Positive
Correlation ($0 \leq r < 0.7$)



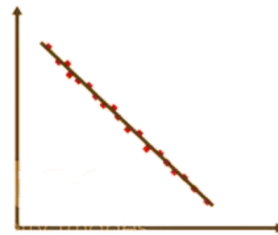
No Correlation ($r \approx 0$)



Weak Negative
Correlation ($-0.7 \leq r < 0$)

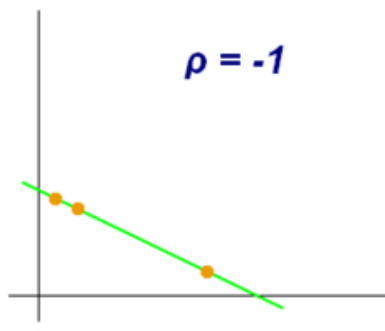


Weak Negative
Correlation ($-0.7 \leq r < 0$)

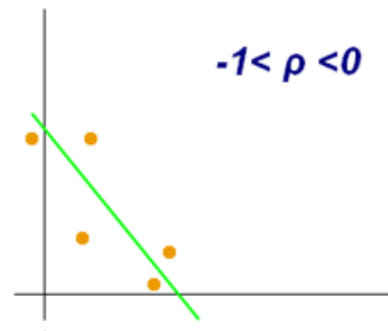


Perfect Negative
Correlation ($r = -1$)

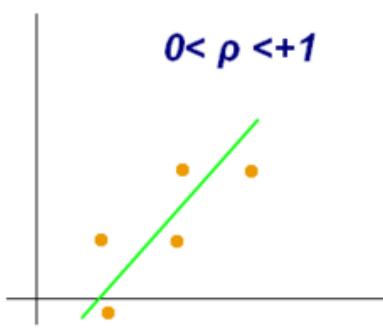
Προσοχή: Στο μεσαίο παραπάνω σχήμα, υπάρχει λάθος! Η συσχέτιση δεν είναι ασθενής αλλά ισχυρή αρνητική συσχέτιση.



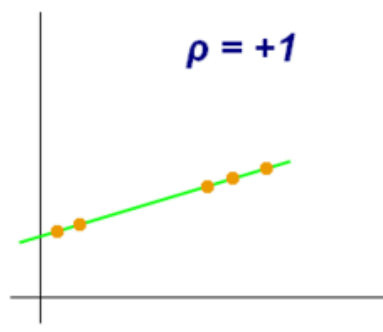
$\rho = -1$



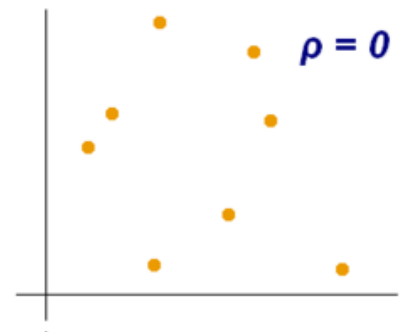
$-1 < \rho < 0$



$0 < \rho < +1$



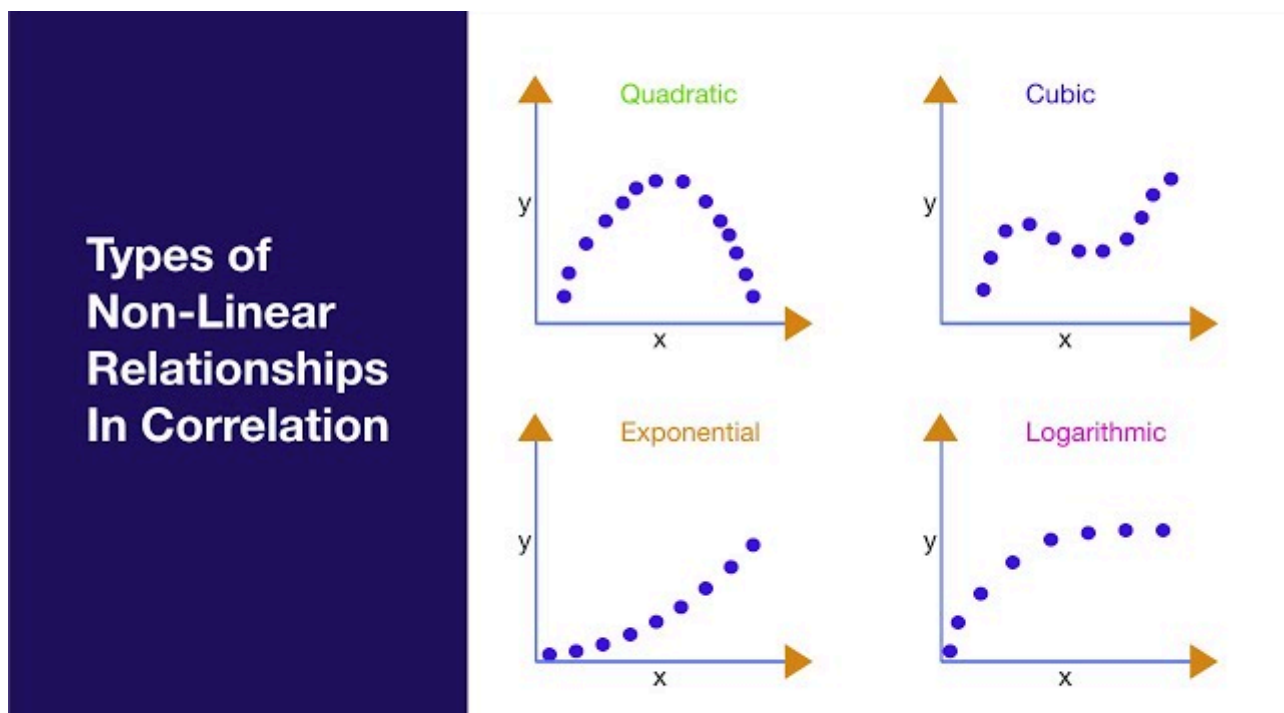
$\rho = +1$



$\rho = 0$

Interval Coefficient	Level of relationship
0.00 – 0.199	Very Weak
0.20 – 0.399	Weak
0.40 – 0.599	Medium
0.60 – 0.799	Strong
0.80 – 1.000	Very Strong

Στο παρακάτω σχήμα απεικονίζονται μη γραμμικές (nonlinear) συσχετίσεις για τις οποίες δεν ισχύει ο συντελεστής συσχέτισης (και οι οποίες είναι εκτός του αντικειμένου αυτού του εισαγωγικού μαθήματος)



8. Πιθανότητες και κατανομές πιθανοτήτων

Μόνο βασικές και χρήσιμες γνώσεις

8.1. Στοιχειώδεις γνώσεις πιθανοτήτων

***** Η ενότητα αυτή (8.1) εκτός ύλης για την πρόοδο 1 *****

Simple Probability

$$\text{Probability} = \frac{\text{Favorable outcomes}}{\text{Total outcomes}}$$



Example:

$$P(\text{red}) = \frac{7}{12}$$

← Number of red marbles
← Total number of marbles (sample space)

$$P(\text{blue}) = \frac{5}{12}$$

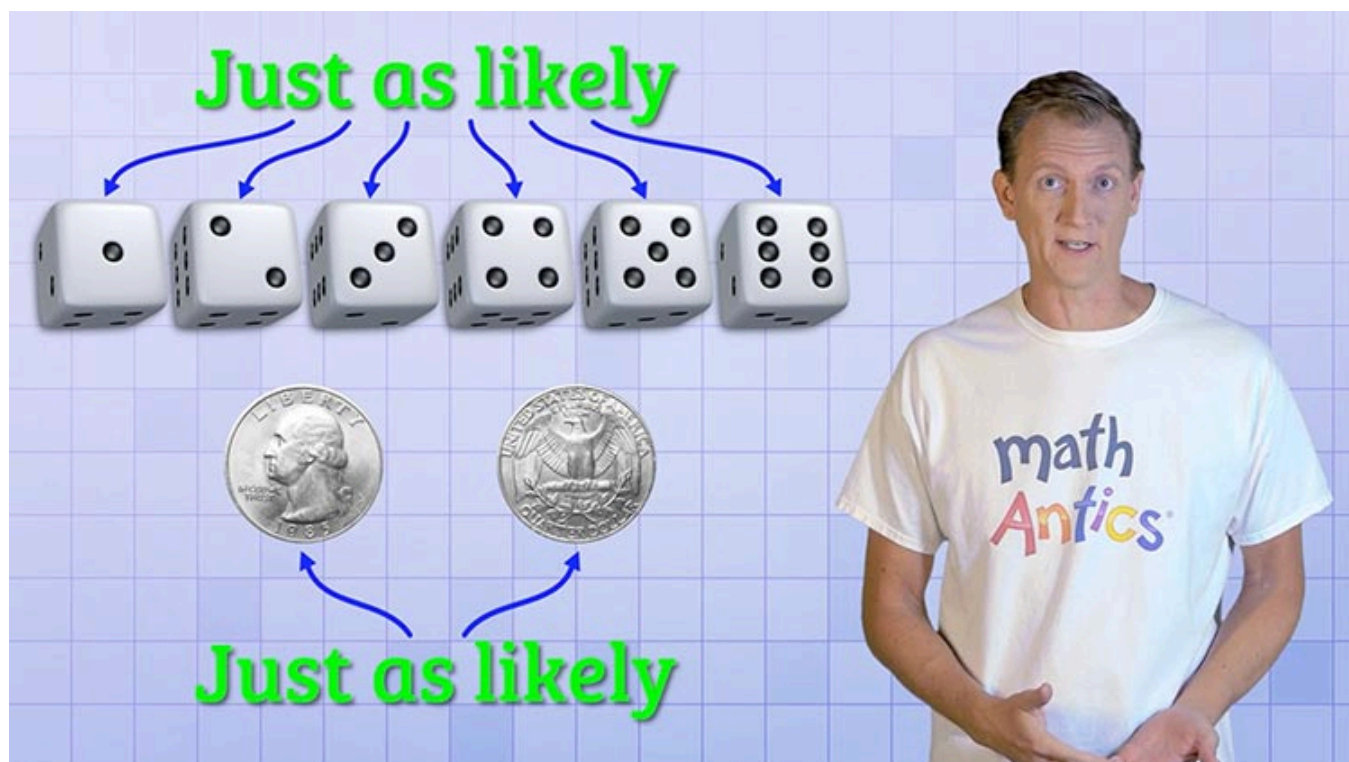
← Number of blue marbles
← Total number of marbles (sample space)

Basic Properties of Probabilities

Property 1: The probability of an event is always between 0 and 1, inclusive.

Property 2: The probability of an event that cannot occur is 0. (An event that cannot occur is called an **impossible event**.)

Property 3: The probability of an event that must occur is 1. (An event that must occur is called a **certain event**.)



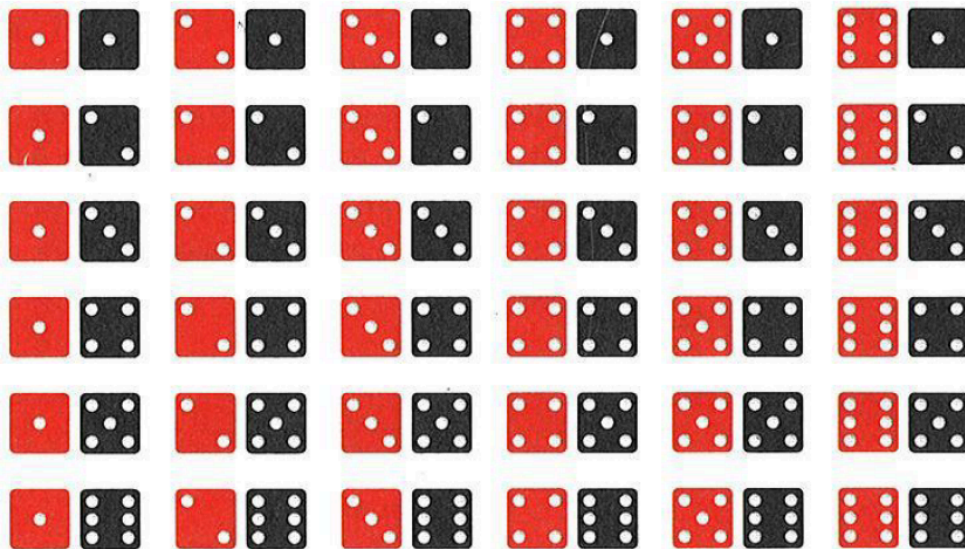
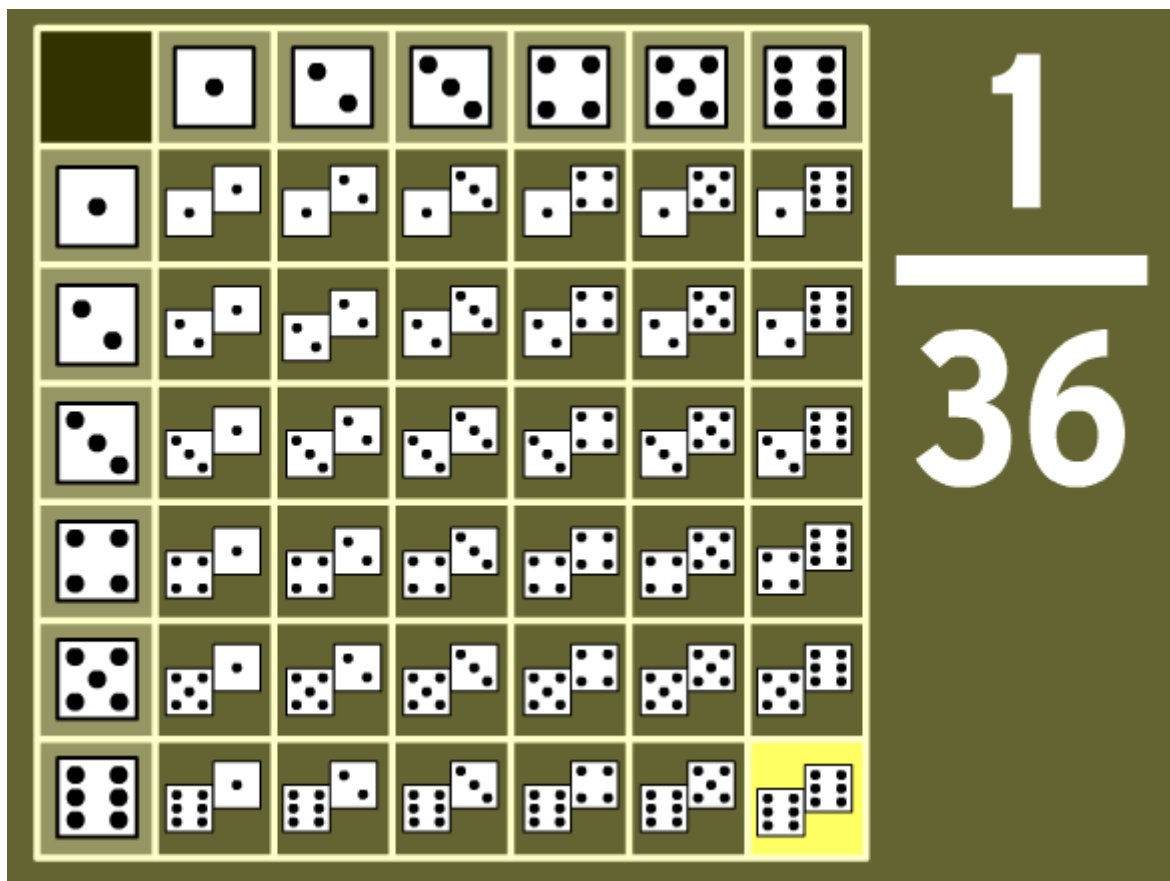
A bag contains ³3 blue marbles, ²2 red marbles and ¹1 green marble. If you were to draw one marble at random, what is the probability of drawing...

1. A blue marble? $\frac{3}{6} = \boxed{\frac{1}{2}}$

2. A red marble? $\frac{2}{6} = \boxed{\frac{1}{3}}$

3. A black marble? $\frac{0}{6} = \boxed{0}$

4. A blue or green marble?



b. Figure 5.1 also shows that doubles can be rolled in six ways. Consequently, the probability of rolling doubles equals $f/N = 6/36 = 0.167$. ♦

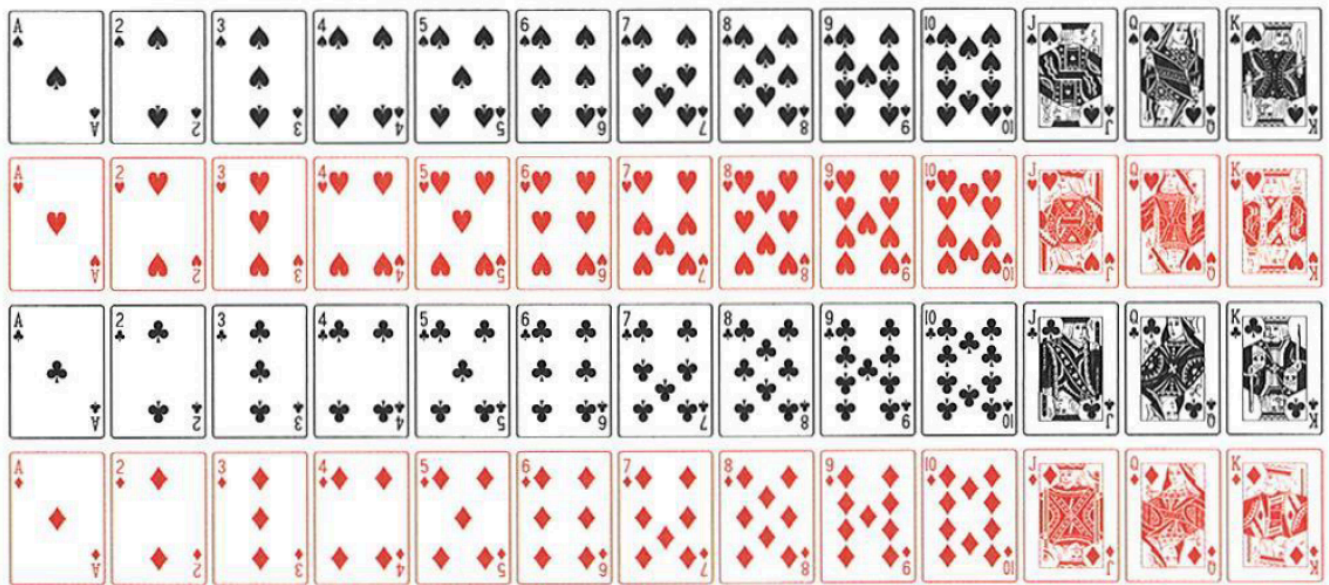
5.12 Nobel Prize Winners. The National Science Foundation collects data on Nobel Prize Laureates in the field of science and the date and location of their award-winning research. A frequency distribution for the number of winners, by country, for the years 1901–1997 are as follows.

Country	Winners
United States	190
United Kingdom	71
Germany	61
France	25
Soviet Union	10
Japan	4
Other Countries	87

Suppose that a recipient of a Nobel Prize in science between 1901–1997 is selected at random. Find the probability that the Nobel Laureate is from

- a. Japan.
- b. either France or Germany.
- c. any country other than the United States.

Playing Cards A deck of playing cards contains 52 cards, as displayed in Fig. 5.3. When we perform the experiment of randomly selecting one card from the deck, one of these 52 cards will be obtained. The collection of all 52 cards—the possible outcomes—is called the **sample space** for this experiment.



Many different events can be associated with this card-selection experiment. Let's consider four:

1. The event that the card selected is the king of hearts.
2. The event that the card selected is a king.
3. The event that the card selected is a heart.
4. The event that the card selected is a face card.

FIGURE 5.5
The event a king is selected

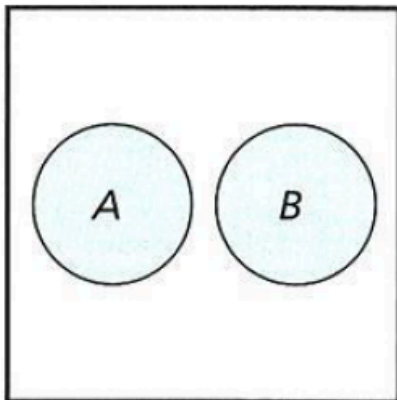
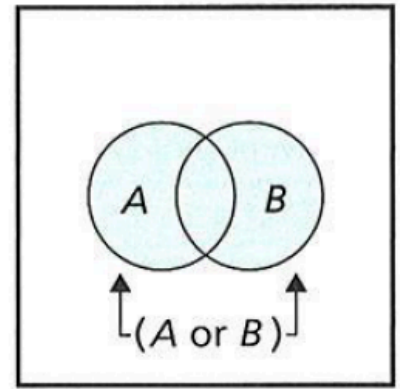
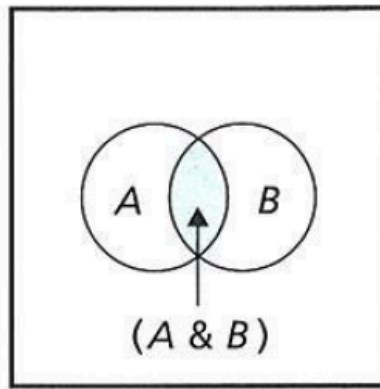
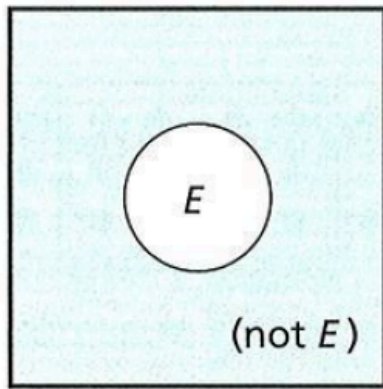


FIGURE 5.6
The event a heart is selected

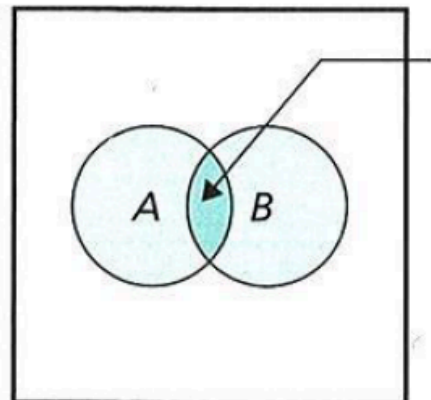


FIGURE 5.7
The event a face card is selected



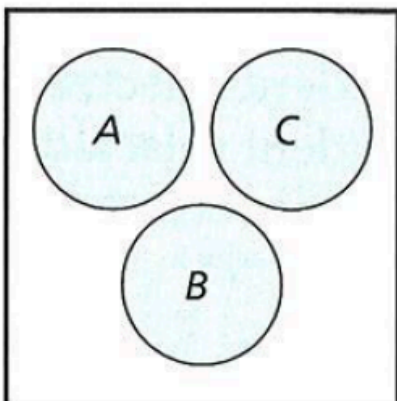


(a)

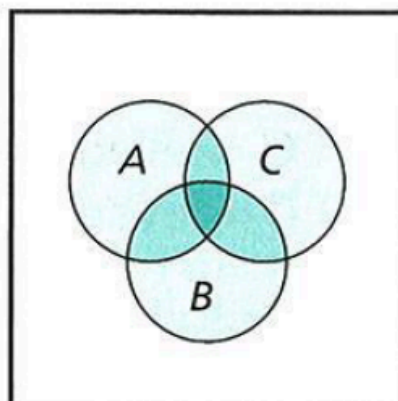


Common
outcomes

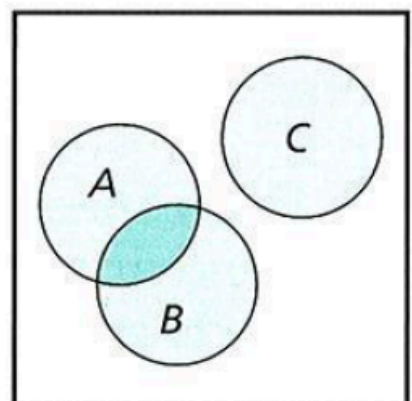
(b)



(a)



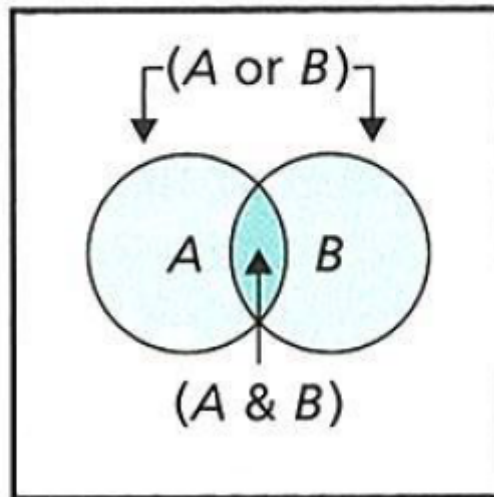
(b)



(c)

FIGURE 5.19

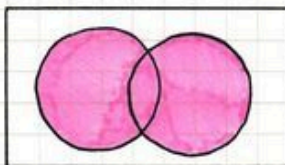
Non-mutually exclusive events



Basic Probability

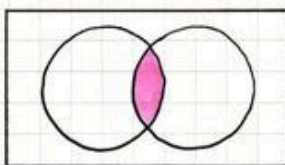
REVISION OF SET NOTATION

① UNION



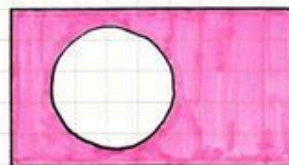
$A \cup B$ EITHER A OR B OR BOTH

② INTERSECTION



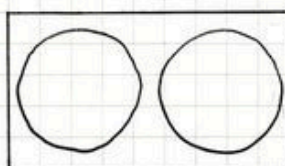
$A \cap B$ IN BOTH A AND B

③ COMPLEMENT



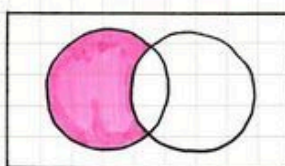
A' NOT IN A

④ MUTUALLY EXCLUSIVE



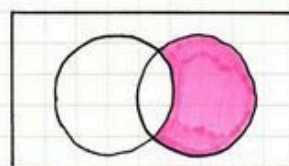
NO INTERSECTION

⑤ $A \cap B'$



IN A BUT NOT IN B

⑥ $A' \cap B$



IN B BUT NOT IN A

ONE WAY TO FIGURE OUT WHICH BIT TO SHADE: SHADE ALL REQUIRED BITS FOR A AND B FIRST, THEN IF IT IS

- UNION ($\cup$), ALL BITS SHADED ARE INCLUDED IN FINAL DIAGRAM
- INTERSECTION ($\cap$), ONLY BITS WITH OVERLAPPING SHADING (FROM BOTH A AND B) ARE INCLUDED.

PROBABILITY FACTS

$$Pr(A) = \frac{\text{NO. FAVOURED OUTCOMES}}{\text{NO. POSSIBLE OUTCOMES}}$$

$$0 \leq Pr(A) \leq 1$$

↑ ↑
IMPOSSIBLE CERTAIN

$$Pr(A') = 1 - Pr(A)$$

IF $Pr(A \cap B) = 0$ THEN A AND B ARE
MUTUALLY EXCLUSIVE

IF $Pr(A \cap B) = Pr(A) \times Pr(B)$ } THEN A AND B
OR IF $Pr(A|B) = Pr(A)$ } ARE INDEPENDENT

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

↗ BOTH EVENTS OCCUR
↘ GIVEN ONE OCCURS

A GIVEN THAT B OCCURS / A IF B OCCURS.

* SAMPLE SPACE GIVEN IN $\{ \}$!

COMBINATIONS

THE NUMBER OF WAYS TO SELECT A GIVEN NO. OF ITEMS
WHEN ORDER DOES NOT MATTER.

IT IS GIVEN BY:

$${}^nC_r \text{ OR } \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

WHEN SELECTING r OBJECTS
FROM A GROUP OF n OBJECTS

REMEMBER THESE!

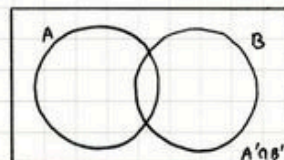
$$\binom{n}{r} = \binom{n}{n-r} \quad \binom{n}{0} = 1 = \binom{n}{n} \quad \binom{n}{1} = n = \binom{n}{n-1}$$

$$\binom{n}{2} = \binom{n}{n-2}$$

DIAGRAMS

	A	A'
B	$Pr(A \cap B)$	$Pr(A' \cap B)$
B'	$Pr(A \cap B')$	$Pr(A' \cap B')$
	$Pr(A)$	$Pr(A')$

KARNAUGH MAP
 $Pr(A \cup B)$



VENN DIAGRAM

Probability Formulas

$$P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \cap B) = P(A/B) \times P(B)$$

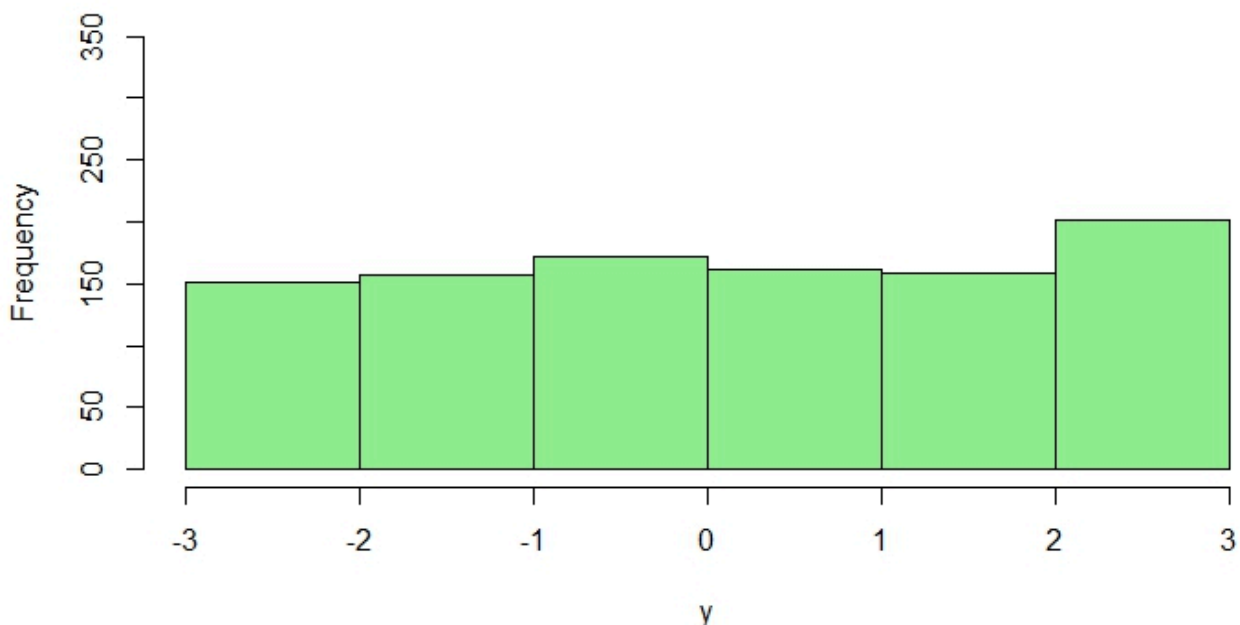
$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)}$$

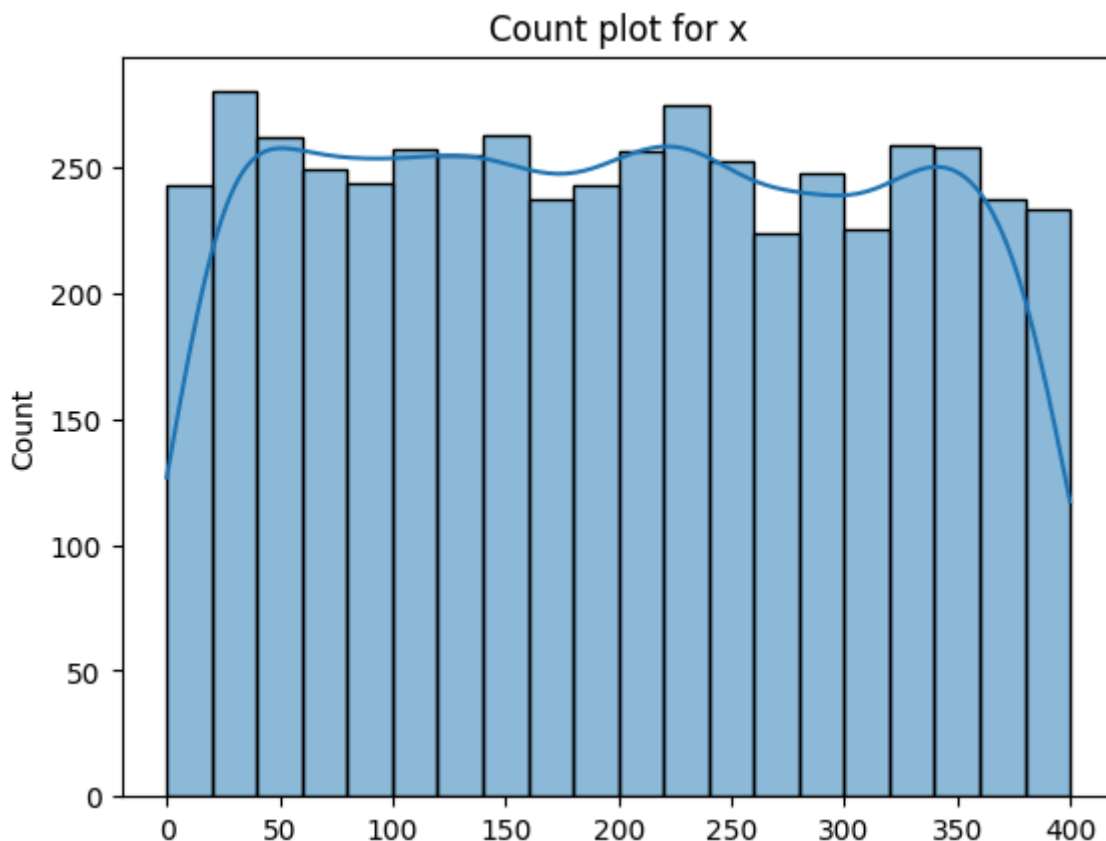
8.2. Κατανομές πιθανοτήτων

***** ΑΠΟ ΕΔΩ ΚΑΙ ΠΕΡΑ ΔΙΑΒΑΖΕΤΕ ΓΙΑ ΤΗΝ ΠΡΟΟΔΟ 2 *****

8.2.1. Ομοιόμορφη κατανομή

UNIFORM DISTRIBUTION





Uniform Distribution Formula



$$F(x) = \frac{1}{(b - a)}$$

$$\text{Mean} = \frac{(a + b)}{2}$$

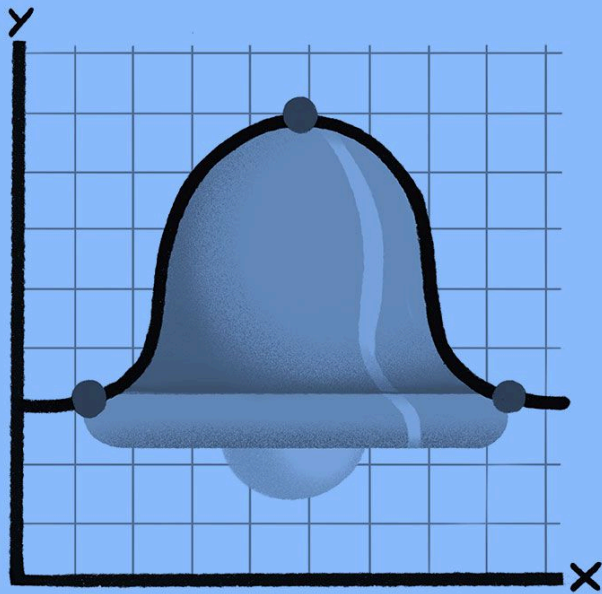
$$\sigma = \sqrt{\frac{(b - a)^2}{12}}$$

8.2.2. Κανονική κατανομή

Η πιο σημαντική κατανομή

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

- Καμπύλη σαν καμπάνα (bell curve)



Bell Curve

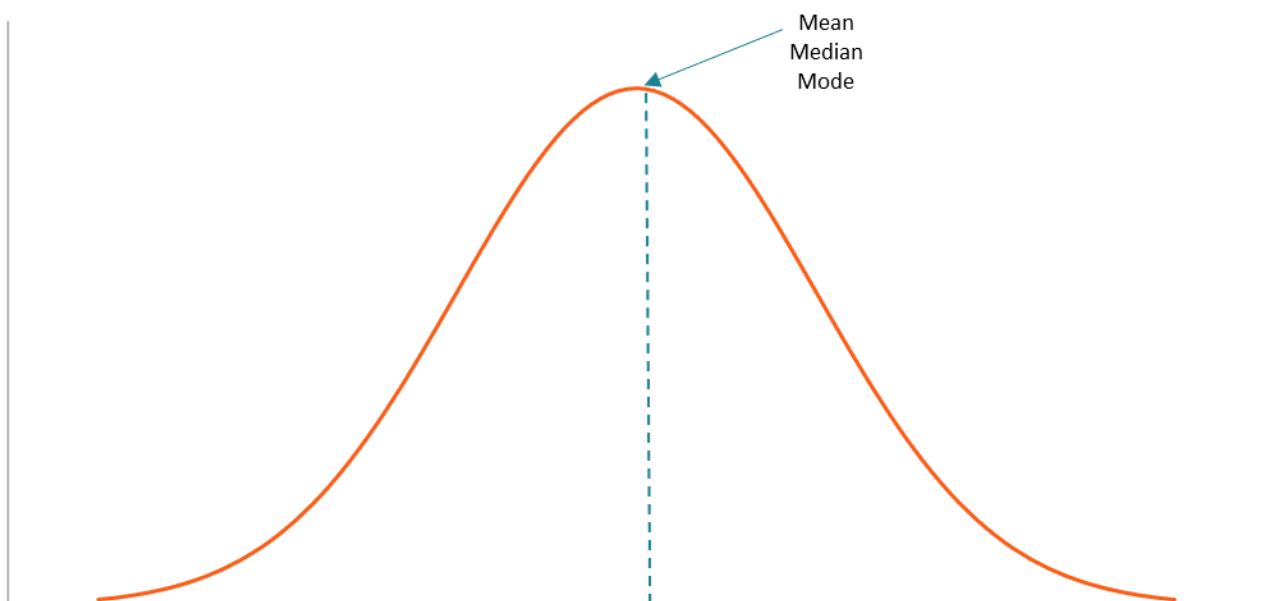
[ˈbel ˈkərv]

A common type of distribution for a variable, also known as the normal distribution.



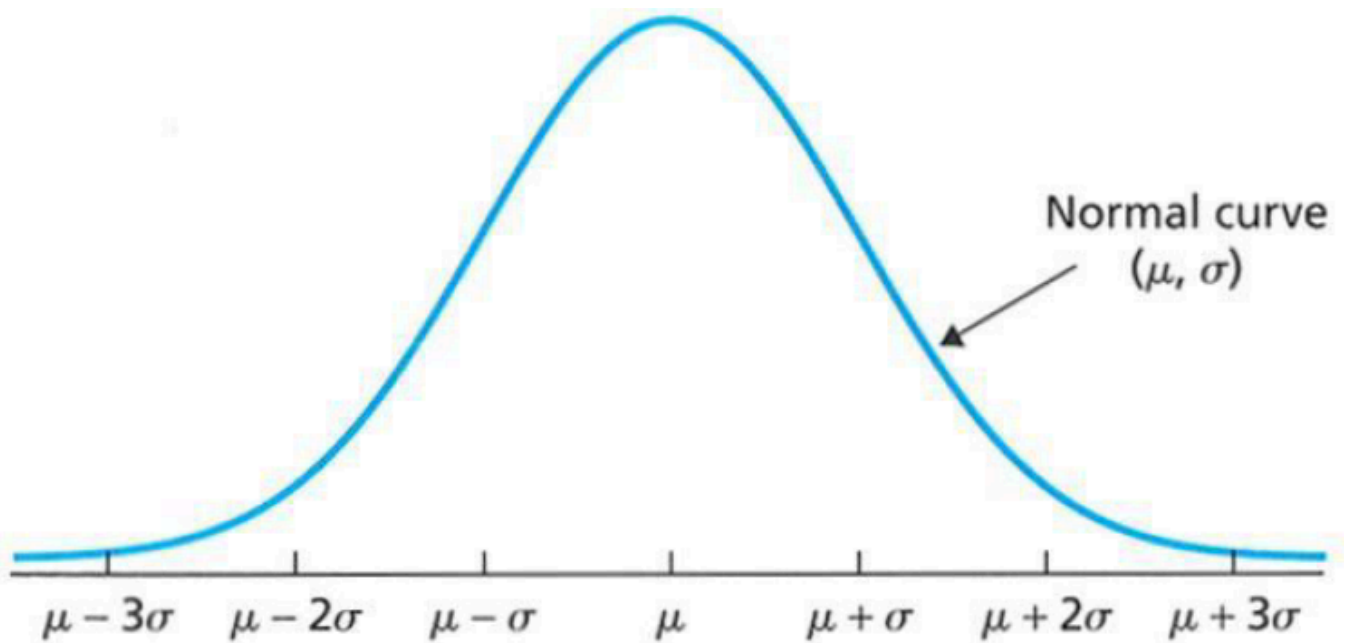
Τυπικά χαρακτηριστικά της κανονικής καμπύλης

- Συμμετρική
 - Μηδενική στρέβλωση (skewness)
- Ελάχιστο = μείον άπειρο ($-\infty$)
- Μέγιστο = συν άπειρο ($+\infty$)

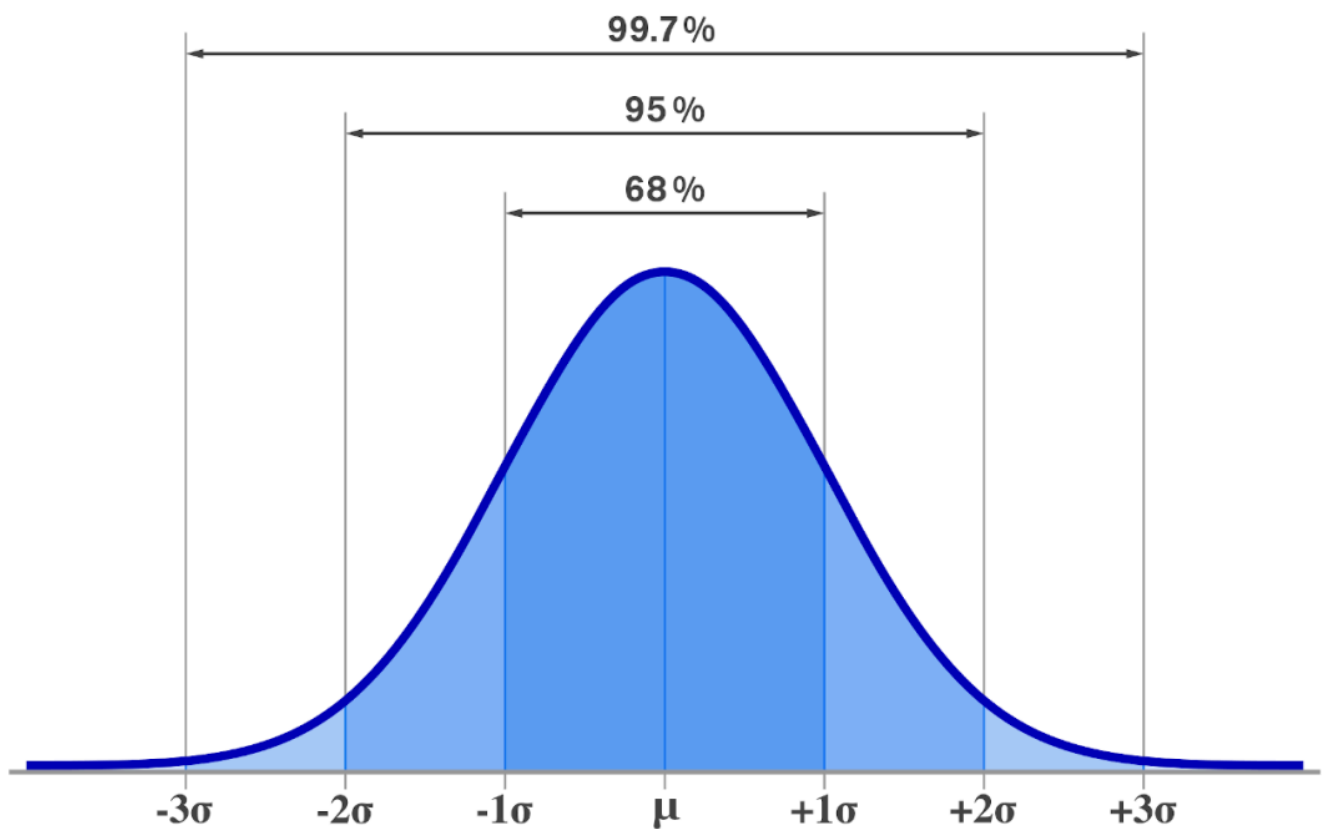


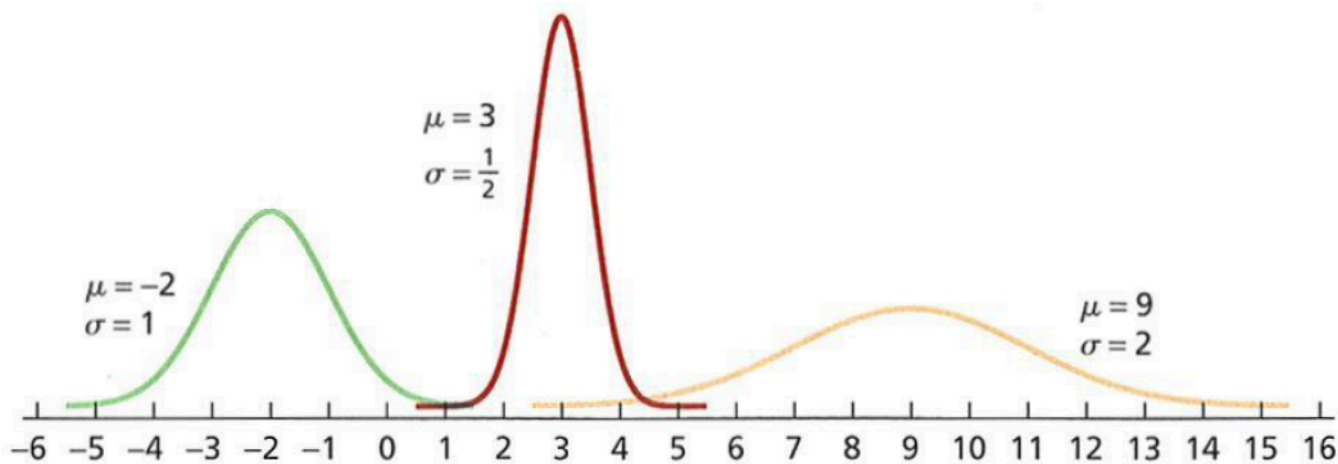
Μια κανονική κατανομή καθορίζεται πλήρως από τη μέση τιμή της, μ , και στην τυπική της απόκλιση, σ

- Και συμβολίζεται $N(\mu, \sigma)$



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



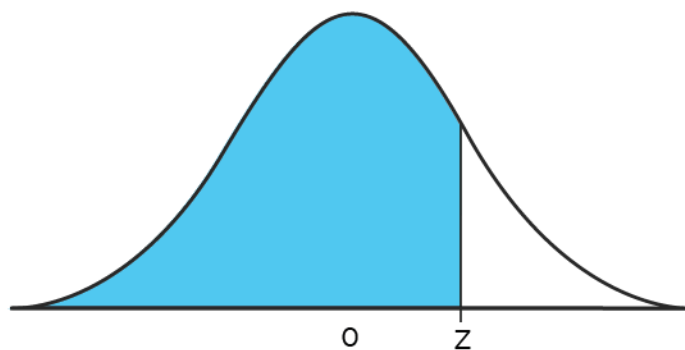


Πίνακες κανονικής κατανομής

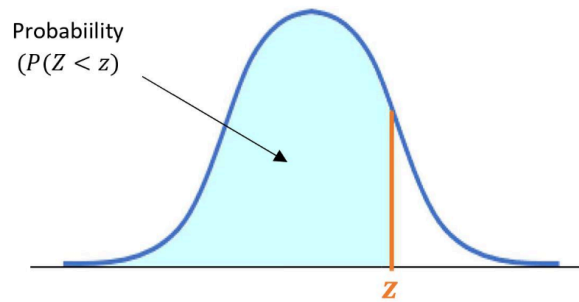
τυποποιημένη ή μοναδιαία κανονική κατανομή

https://onlinestatbook.com/lms/calculators/normal_dist.html

Table of Standard Normal Probabilities for Positive 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

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

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.

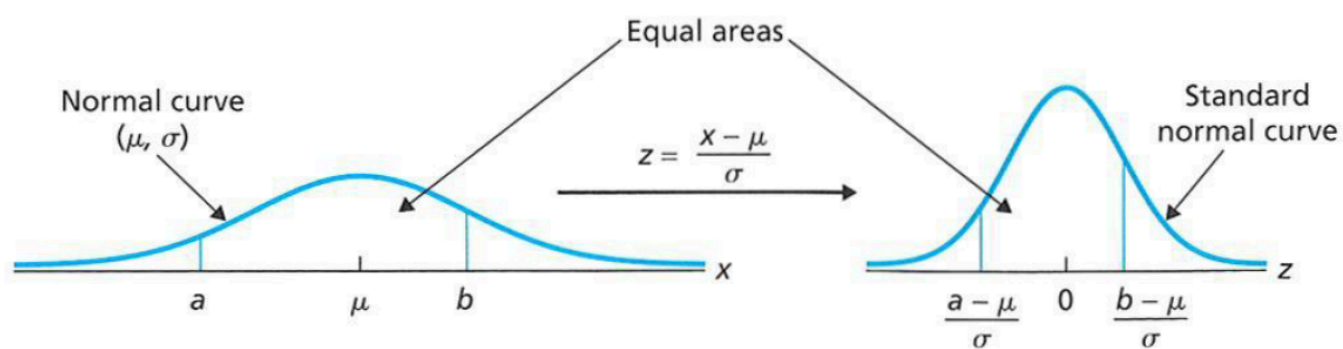
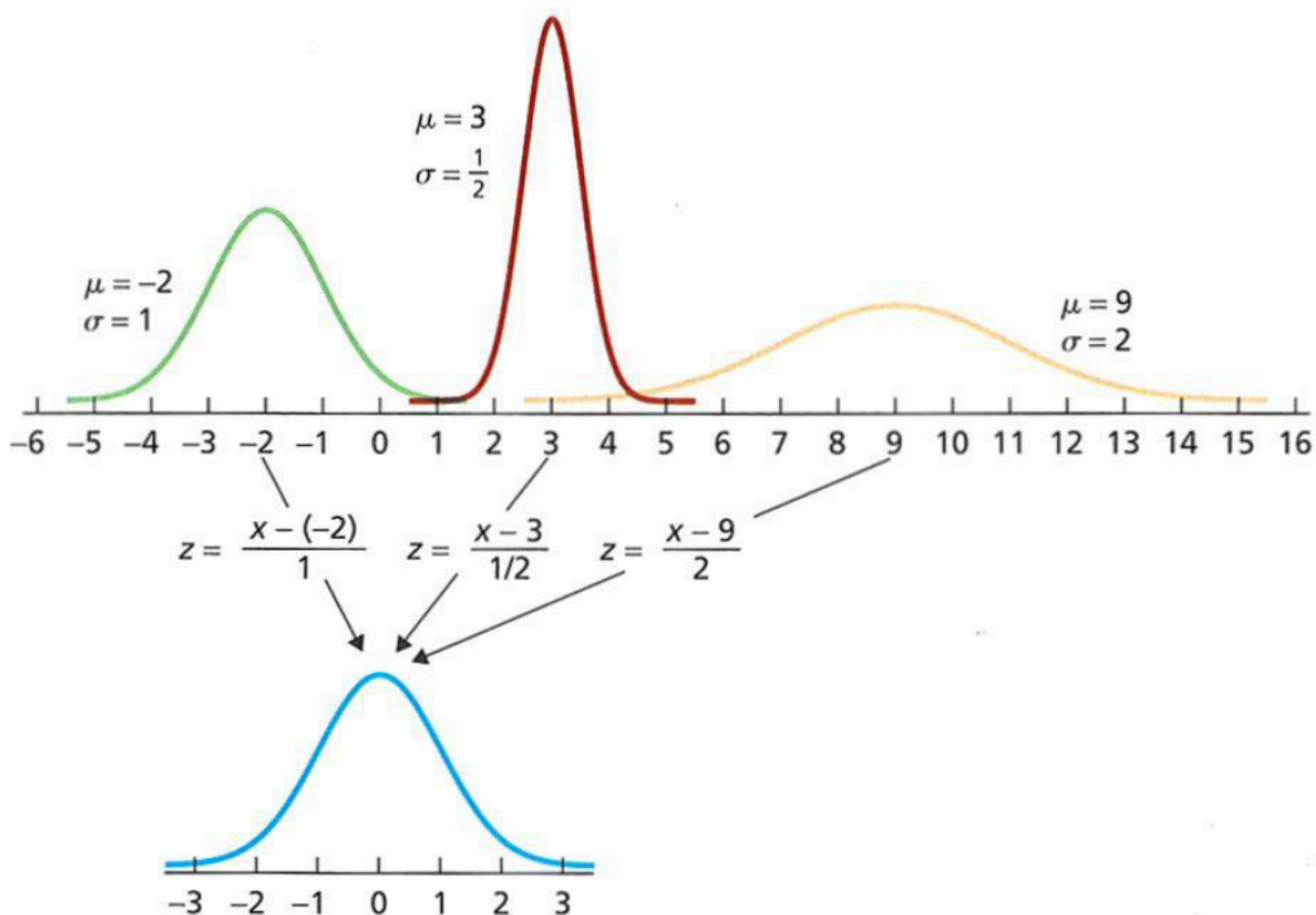
Assuming a normal distribution, in which race did I perform the best compared to the other participants?

$$z = \frac{x - \mu}{\sigma} \quad \text{200m}$$

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

$$z = \frac{x - \mu}{\sigma} \quad \text{500m}$$

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



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 (1×10) points, or 10 points below the mean of 50; that is, $X = 40$.

Εμπειρικός κανόνας

(1) Διάστημα ($\mu - \sigma, \mu + \sigma$)

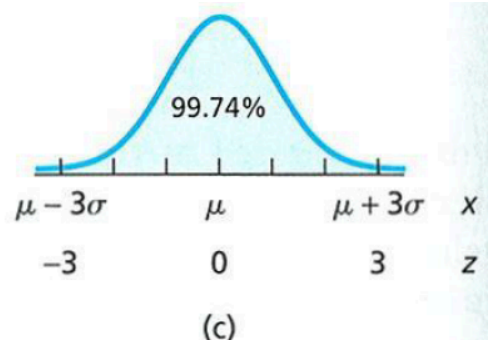
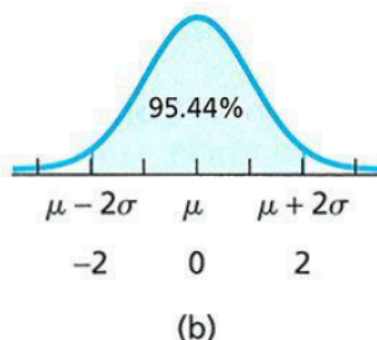
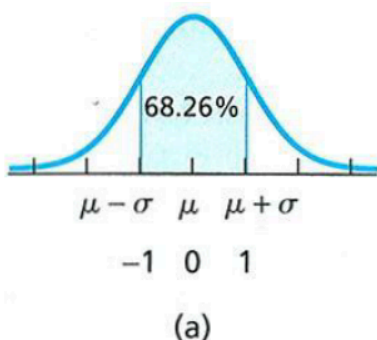
- 68.26%
- (λίγο παραπάνω από τα δύο τρίτα, περίπου 68%)

(2) Διάστημα ($\mu - 2\sigma, \mu + 2\sigma$)

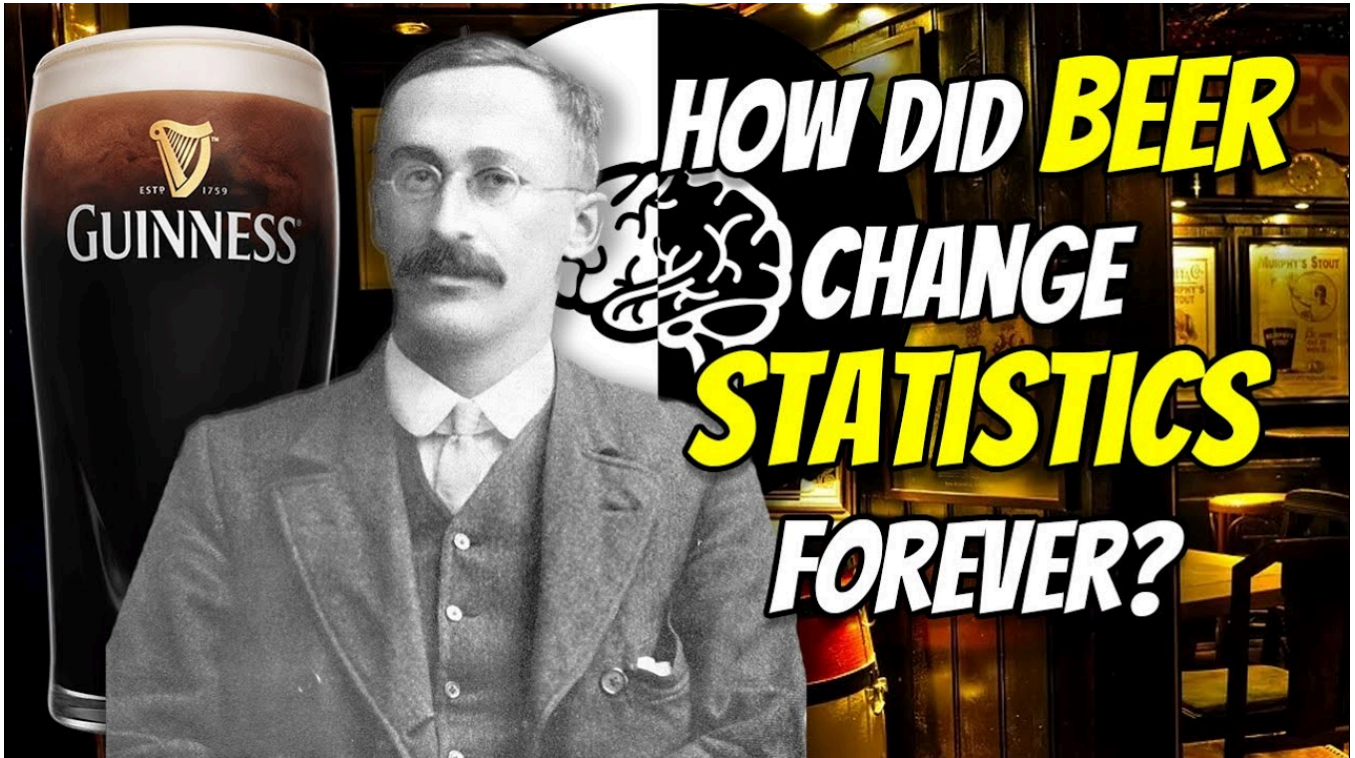
- 95.44%
- (περίπου 95%)

(3) Διάστημα ($\mu - 3\sigma, \mu + 3\sigma$)

- 99.74%
- (περίπου όλα)



8.2.3. Κατανομή t (του σπουδαστή)

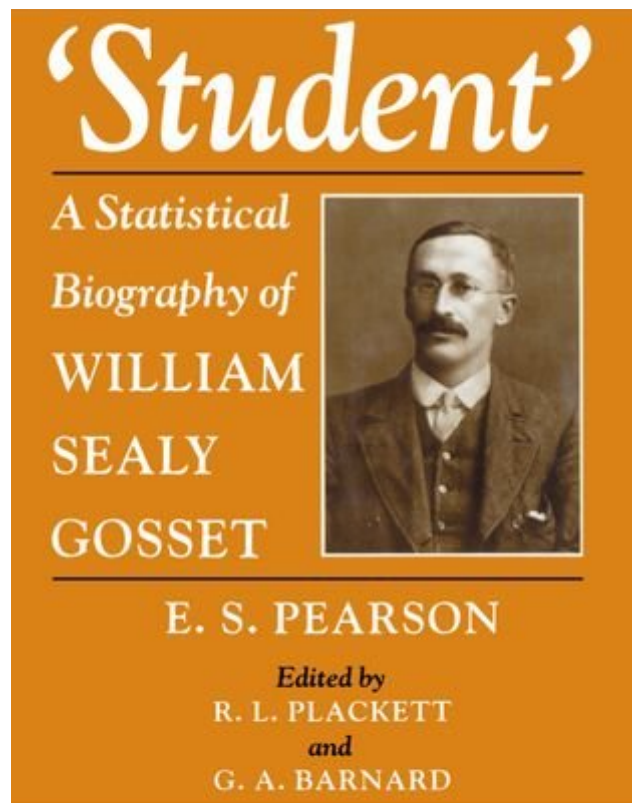


William S. Gosset

Who invented t -test?

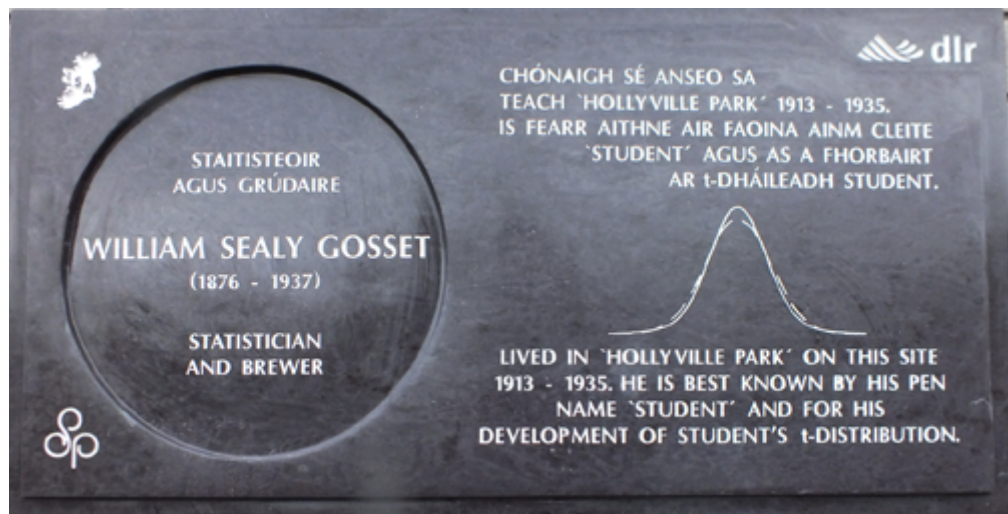
The t -test is 107 years old. The t statistic was introduced by William Sealy Gosset. Gosset published the t test in Biometrika in 1908, but was forced to use a pen name by his employer who regarded the fact that they were using statistics as a trade secret. Hence, the name **Student's t -test**.





"The new plaque honoring statistician William Sealy Gosset is displayed at the Saint Patrick's Boys National Primary School in Dublin, Ireland"

(<https://magazine.amstat.org/blog/2013/02/01/gossetplaque/>)





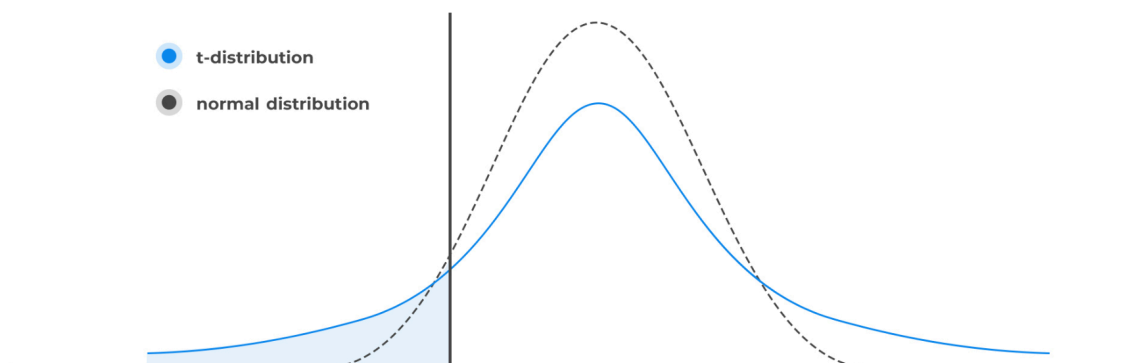
Ronald A. Fisher



βαθμοί ελευθερίας (degrees of freedom)



T-distribution vs Normal Distribution



Online demo της σχέσης ανάμεσα στην κανονική κατανομή και την κατανομή t

- <https://www.desmos.com/calculator/xm56tvvalh>

Basic Properties of t -Curves

Property 1: The total area under a t -curve equals 1.

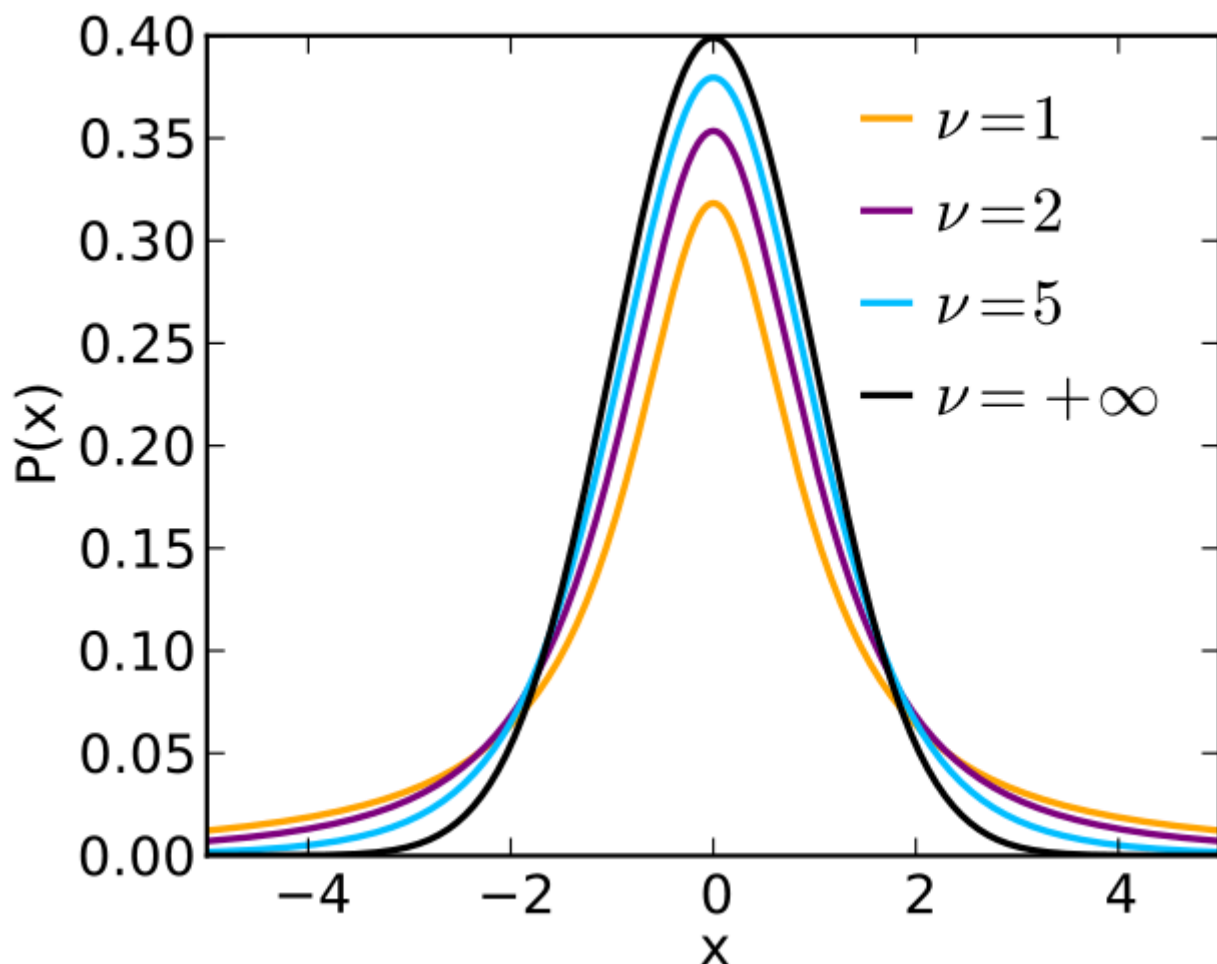
Property 2: A t -curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.

Property 3: A t -curve is symmetric about 0.

Property 4: As the number of degrees of freedom becomes larger, t -curves look increasingly like the standard normal curve.

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)



8.2.4. Κατανομή F

***** Εκτός ύλης *****

πηλίκο δύο κατανομών t

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

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

8.2.5. Κατανομή χ^2

Ιδιότητες της κατανομής χ^2

- Χαρακτηρίζεται από βαθμούς ελευθερίας (degrees of freedom)
 - Όσο μεγαλώνουν οι βαθμοί ελευθερίας, τόσο το σχήμα της κατανομής χ^2 προσεγγίζει την κανονική κατανομή
 - Όπως και η κατανομή t
- Δεν παίρνει αρνητικές τιμές
 - Δηλαδή παίρνει τιμές μεγαλύτερες ή ίσες του μηδενός

- Έχει ως ελάχιστο το μηδέν
 - Σε αυτό διαφέρει ουσιωδώς από την κανονική κατανομή όσο πολλοί και να είναι οι βαθμοί ελευθερίας
- Έχει ως μέγιστο το $+\infty$ (συν άπειρο)
- Έχει θετική στρέβλωση

Basic Properties of χ^2 -Curves

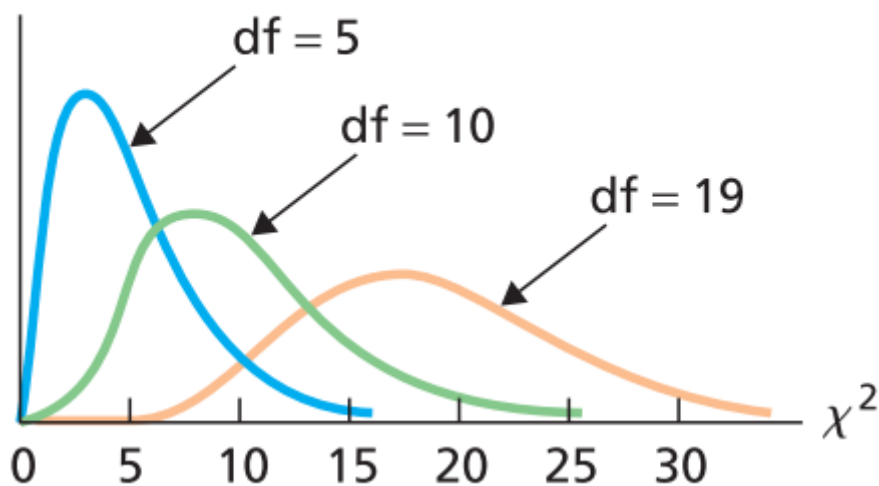
Property 1: The total area under a χ^2 -curve equals 1.

Property 2: A χ^2 -curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis.

Property 3: A χ^2 -curve is right skewed.

Property 4: As the number of degrees of freedom becomes larger, χ^2 -curves look increasingly like normal curves.

χ^2 -curves for df = 5, 10, and 19



πίνακες κατανομής χ^2

8.2.6. Κατανομή βήτα

***** Εκτός ύλης *****

ελάχιστο, κορυφή, μέγιστο

χρήση στον χρονικό και κοστολογικό προγραμματισμό έργων (project management)

9. Δειγματοληπτική κατανομή

Εξετάζοντας τη δειγματοληπτική κατανομή, περνάμε από την Περιγραφική Στατιστική στην Επαγωγική Στατιστική

- Με τις μεθόδους της επαγωγικής στατιστικής αναλύουμε ένα δείγμα (sample) με στόχο να γενικεύσουμε τα ευρήματά μας στον πληθυσμό (population) από τον οποίο προέρχεται το δείγμα.

9.1. Δείγματα και δειγματοληπτικό σφάλμα

Όταν αναλύουμε ένα δείγμα, θα υπάρχει πάντα δειγματοληπτικό (ή δειγματικό) σφάλμα (sampling error)

- Γιατί ένα δείγμα, όσον καλό και να είναι, θα αποτελεί πάντα ένα πολύ μικρό κομμάτι του πληθυσμού
- Επομένως τα δειγματικά μεγέθη (sample statistics) θα διαφέρουν λίγο ή πολύ από τις πληθυσμιακές παραμέτρους (population parameters)

Ο μόνος τρόπος να μην υπάρχει σφάλμα στα αποτελέσματα της ανάλυσής μας, είναι να κάνουμε απογραφή πληθυσμού (census)

- Αυτό όμως θα κόστιζε πολύ και ως εκ τούτου δεν γίνεται σχεδόν ποτέ

Ακόμα και ένα δείγμα, π.χ. 60 χιλιάδων νοικοκυριών, θα έδινε μέσο εισόδημα νοικοκυριού που θα διέφερε (ίσως όχι πολύ) από τον πραγματικό μέσο όρο του εισοδήματος για τον συνολικό πληθυσμό

- Το ερώτημα λοιπόν που πρέπει να απαντήσουμε είναι πόσο ακριβής είναι πιθανό να είναι η εκτίμηση μιας πληθυσμιακής παραμέτρου από ένα δειγματικό μέγεθος

Την απάντηση σε αυτό το εξαιρετικά σημαντικό ερώτημα δίνει η

δειγματοληπτική κατανομή (sampling distribution), που λέγεται και δειγματική κατανομή

- Όλα τα στατιστικά μεγέθη (μέση τιμή, τυπική απόκλιση, κλπ.) χαρακτηρίζονται από τη δική τους δειγματοληπτική κατανομή
- Εμείς σε αυτή την ενότητα θα εξετάσουμε τη δειγματοληπτική κατανομή του δειγματικού μέσου όρου (sampling distribution of the sample mean), δηλαδή του $\bar{x}$
- Μιλάμε για δείγματα του ίδιου μεγέθους (που είναι και συγκρίσιμα)

Αντιλαμβανόμαστε ότι η μέση τιμή ενός δείγματος θα διαφέρει από το ένα δείγμα (συγκεκριμένου μεγέθους) στο άλλο (του ίδιου μεγέθους)

- Επομένως, η μέση τιμή αυτών των δειγμάτων, $\bar{x}$, θα είναι σαν μια τυχαία μεταβλητή (random variable), της οποίας η τιμή θα αλλάζει από δείγμα σε δείγμα

Ξεκινάμε εξετάζοντας ένα πολύ μικρό πληθυσμό (για εκπαιδευτικούς σκοπούς), που αποτελείται από τους 5 βασικούς παίκτες μιας ομάδας μπάσκετ

- Η μόνη μεταβλητή (variable) που εξετάζουμε είναι το ύψος τους, το οποίο μετριέται σε ίντσες (μια ίντσα = 2.54 εκατοστά)
- Θα συμβολίζουμε το μεταβλητή ύψος με x (όπως συνήθως κάνουμε στη στατιστική)

TABLE 7.1
Heights, in inches, of the five
starting players

Player	A	B	C	D	E
Height	76	78	79	81	86

Η μέση τιμή του ύψους στο μικρό αυτό πληθυσμό είναι

$$\mu = \frac{\sum_{i=1}^5 x_i}{n} = \frac{76+78+79+81+86}{5} = \frac{400}{5} = 80$$

Θα εξετάσουμε τώρα δείγματα μεγέθους 2 (και αυτό πολύ μικρό ως μέγεθος, επίσης για εκπαιδευτικούς σκοπούς)

- Αγνοήστε το γεγονός ότι ένα δείγμα μεγέθους 2 δεν είναι τάξεις μεγέθους μικρότερο του μεγέθους του πληθυσμού (κάτι που κανονικά πρέπει να συμβαίνει)

Στον παρακάτω πίνακα απεικονίζονται όλα τα δυνατά δείγματα (που είναι 10) μεγέθους 2 ($n=2$) από τον παραπάνω πληθυσμό (του οποίου το μέγεθος είναι $N=5$)

- Στη δεξιά στήλη του πίνακα δείχνουμε τη μέση τιμή κάθε δείγματος ($\bar{x}$)

TABLE 7.2
Possible samples and sample means for
samples of size 2

Sample	Heights	$\bar{x}$
A, B	76, 78	77.0
A, C	76, 79	77.5
A, D	76, 81	78.5
A, E	76, 86	81.0
B, C	78, 79	78.5
B, D	78, 81	79.5
B, E	78, 86	82.0
C, D	79, 81	80.0
C, E	79, 86	82.5
D, E	81, 86	83.5

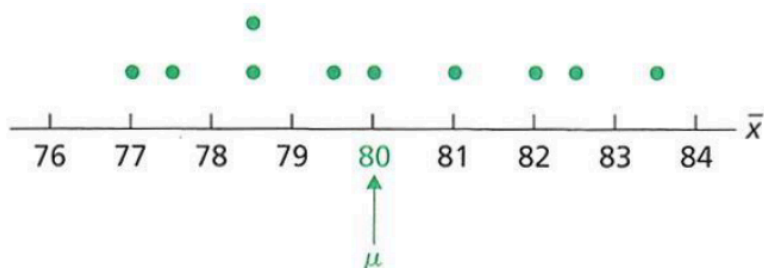
Παρατηρούμε ότι

- Κάποια δείγματα, όπως το (C, D) έχουν μέσο όρο ($\bar{x} = 80$) ακριβώς ίδιο με αυτό το πληθυσμού ($\mu = 80$)
- Κάποια άλλα δείγματα, όπως το (D, E), έχουν μέσο όρο ($\bar{x} = 83.5$) που απέχει αρκετά από το μέσο όρο του πληθυσμού ($\mu = 80$)
- Σε κάθε περίπτωση άλλη από αυτό το εκπαιδευτικό παράδειγμα, δεν θα ξέρουμε αν το δείγμα που έχουμε επιλέξει έχει μέσο όρο κοντά ή μακριά από το μέσο όρο του πληθυσμού
 - Τον οποίο μέσο όρο του πληθυσμού επίσης δεν θα τον γνωρίζουμε!

Παρακάτω απεικονίζουμε τους μέσους όρους των 10 αυτών δειγμάτων μεγέθους 2 σε ένα κουκιδόγραμμα (dot plot)

FIGURE 7.1

Dotplot for the sampling distribution of the sample mean for samples of size 2 ($n = 2$)



Παρατηρούμε το εξής ενδιαφέρον

- Ο μέσος όρος των μέσων όρων των 10 δειγμάτων (δηλαδή όλων των δειγμάτων μεγέθους 2) είναι ίσος με τον μέσο όρο του πληθυσμού

$$\bar{x} = \frac{77+77.5+78.5+81+78.5+79.5+82+80+82.5+83.5}{10} = 80$$

δηλαδή ισχύει το εξής

$$\mu_{\bar{x}} = \mu$$

κάτι που ισχύει για όλα τα μεγέθη δειγμάτων

Επί τη ευκαιρία, για να υπολογίσουμε τον συνολικό αριθμό των δειγμάτων συγκεκριμένου μεγέθους, χρησιμοποιούμε τον παρακάτω τύπο των συνδυασμών (combinations) χωρίς επανάληψη (repetition)

- Το χωρίς επανάληψη, σημαίνει ότι δεν μπορεί να έχουμε ένα δείγμα που να έχει δύο φορές τον ίδιο παίκτη, π.χ. (A, A)
- Επίσης, οι συνδυασμοί σημαίνουν ότι η σειρά με την οποία θα εμφανίζονται οι παίκτες σε ένα δείγμα δεν θα έχει σημασία
 - Π.χ. το δείγμα (A, B) είναι ίδιο με το δείγμα (B, A)

$$C(n, k) = \frac{n!}{(n-k)!k!} = \frac{1 \times 2 \times 3 \times \dots \times n}{[1 \times 2 \times 3 \times \dots (n-k)] \times (1 \times 2 \times 3 \times \dots \times k)}$$

όπου n συμβολίζει το μέγεθος του πληθυσμού και k το μέγεθος του δείγματος

ΠΡΟΣΟΧΗ: Αλλού στις σημειώσεις αυτές έχουμε μάθει να συμβολίζουμε με N το μέγεθος του πληθυσμού και με n το μέγεθος του δείγματος

Ο τύπος φαίνεται λιγάκι περίπλοκος, αλλά η εφαρμογή του είναι εύκολη

- Προσοχή χρειάζεται ο υπολογισμός του παραγοντικού (factorial), που ορίζεται ως εξής για οποιονδήποτε θετικό ακέραιο αριθμό n

$$n! = 1 \times 2 \times 3 \times \dots \times n$$

- Παράδειγμα

$$5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$$

- Ειδικά για το μηδέν, το παραγοντικό ορίζεται ως εξής

$$0! = 1$$

Υπολογίζουμε λοιπόν τα 10 δείγματα μεγέθους 2 ως εξής

$$C(5, 2) = \frac{5!}{(5-2)! \times 2!} = \frac{5!}{3! \times 2!} = \frac{1 \times 2 \times 3 \times 4 \times 5}{(1 \times 2 \times 3) \times (1 \times 2)} = \frac{120}{12} = 10$$

Εάν η σειρά ήταν σημαντική, δηλαδή το δείγμα (A, C) θεωρείτο διαφορετικό από το δείγμα (C, A), θα έπρεπε να κάνουμε χρήση άλλου τύπου, αυτού για τις αποκαλούμενες μεταθέσεις (permutations) χωρίς επανάληψη

- Δηλαδή, ο όρος συνδυασμοί αναφέρεται σε μη διατεταγμένα δείγματα ενώ ο όρος μεταθέσεις σε διατεταγμένα

Ας πάρουμε τώρα δείγματα μεγαλύτερου μεγέθους, να δούμε τι συμβαίνει με τη δειγματοληπτική κατανομή.

Όλα τα δείγματα μεγέθους 4 απεικονίζονται στον παρακάτω πίνακα, με τη μέση τιμή κάθε δείγματος να είναι στη δεξιά στήλη

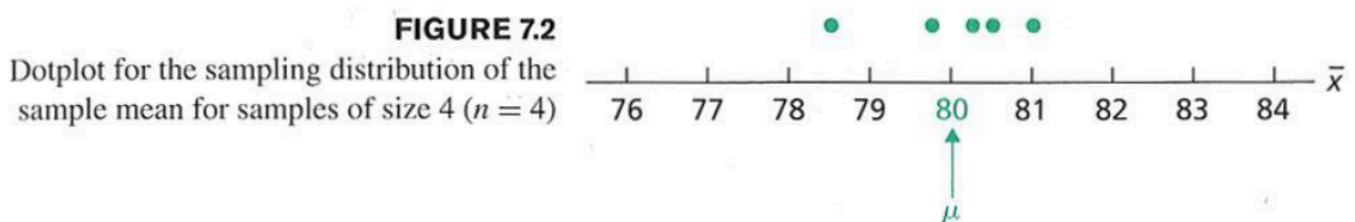
TABLE 7.3
Possible samples and sample means for
samples of size 4

Sample	Heights	$\bar{x}$
A, B, C, D	76, 78, 79, 81	78.50
A, B, C, E	76, 78, 79, 86	79.75
A, B, D, E	76, 78, 81, 86	80.25
A, C, D, E	76, 79, 81, 86	80.50
B, C, D, E	78, 79, 81, 86	81.00

Για να δούμε μήπως ξεχάσαμε κανένα δείγμα – ο συνολικός αριθμός δειγμάτων είναι όντως 5

$$C(5, 4) = \frac{5!}{(5-4)! \times 4!} = \frac{5!}{1! \times 4!} = \frac{1 \times 2 \times 3 \times 4 \times 5}{(1) \times (1 \times 2 \times 3 \times 4)} = \frac{120}{24} = 5$$

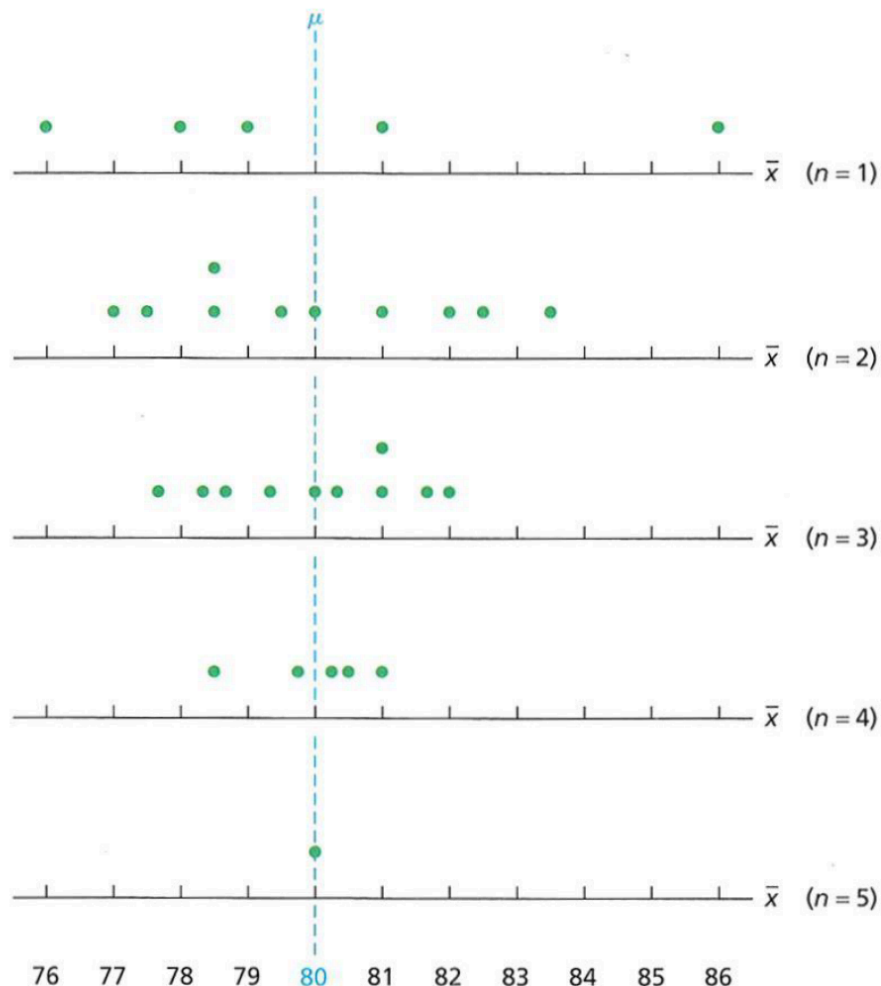
Απεικόνιση των μέσων όρων αυτών των 5 δειγμάτων σε dotplot



Εάν υπολογίζαμε τους μέσους όρους όλων των δειγμάτων μεγέθους 1 έως και 5 και τους απεικονίζαμε σε dot plots, θα παίρναμε το παρακάτω διάγραμμα

FIGURE 7.3

Dotplots for the sampling distributions of the sample mean for samples of sizes 1, 2, 3, 4, and 5



Κάθε δείγμα είναι σαν να φέρει τα χαρακτηριστικά (το "DNA") της "οικογενείας" στην οποία ανήκει

- Που καθορίζεται από το μέγεθός του

Παρατηρούμε ότι όσο μεγαλώνει το μέγεθος του δείγματος, τόσο πιο κοντά στον μέσο όρο του πληθυσμού μαζεύονται οι μέσοι όροι των δειγμάτων!

- Βέβαια, στην πραγματικότητα πάντα διαλέγουμε ένα μόνο δείγμα, για τον μέσο όρο του οποίου δεν ξέρουμε πόσο κοντά στον μέσο όρο του πληθυσμού είναι

Sample size n	No. possible samples	No. within 1" of μ	% within 1" of μ	No. within 0.5" of μ	% within 0.5" of μ
1	5	2	40%	0	0%
2	10	3	30%	2	20%
3	10	5	50%	2	20%
4	5	4	80%	3	60%
5	1	1	100%	1	100%

9.2. Σχήμα και χαρακτηριστικά της δειγματοληπτικής κατανομής

Από το προηγούμενο παράδειγμα, παρατηρούμε ότι ο μέσος όρος των μέσων όρων όλων των δειγμάτων συγκεκριμένου μεγέθους, ισούται με τον μέσο όρο του πληθυσμού

$$\mu_{\bar{x}} = \mu$$

Αυτό είναι μια θεωρητική ιδιότητα που ισχύει για όλες τις δειγματοληπτικές κατανομές του μέσου όρου

- Δηλαδή ισχύει για όλα τα μεγέθη δειγμάτων

Η τυπική απόκλιση της δειγματοληπτικής κατανομής του μέσου όρου ονομάζεται **τυπικό σφάλμα** (standard error) και υπολογίζεται με τον παρακάτω τύπο

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Για να επαληθεύσουμε την ισχύ αυτού του τύπου επιστρέφοντας στο προηγούμενο παράδειγμα των 5 παικτών

TABLE 7.1

Heights, in inches, of the five starting players

Player	A	B	C	D	E
Height	76	78	79	81	86

Θυμόμαστε ότι η μέση τιμή αυτού του πληθυσμού είναι

$$\mu = \frac{\sum_{i=1}^5 x_i}{n} = \frac{76+78+79+81+86}{5} = \frac{400}{5} = 80$$

Η τυπική απόκλιση του πληθυσμού είναι

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} \\ &= \sqrt{\frac{(76-80)^2 + (78-80)^2 + (79-80)^2 + (81-80)^2 + (86-80)^2}{5}}\end{aligned}$$

$$\begin{aligned}
&= \sqrt{\frac{(-4)^2 + (-2)^2 + (-1)^2 + (1)^2 + (6)^2}{5}} \\
&= \sqrt{\frac{16 + 4 + 1 + 1 + 36}{5}} \\
&= \sqrt{\frac{58}{5}} = \sqrt{11.6} = 3.406
\end{aligned}$$

Υπενθυμίζω ότι στον παραπάνω τύπο διαιρούμε με n και όχι με $(n-1)$ γιατί μιλάμε για την τυπική απόκλιση του πληθυσμού (και όχι κάποιου δείγματος)

Ας υπολογίσουμε τώρα την τυπική απόκλιση της δειγματοληπτικής κατανομής του μέσου όρου των δειγμάτων μεγέθους 2 με βάση την παραπάνω τιμή της τυπικής απόκλισης του πληθυσμού και τον τύπο του τυπικού σφάλματος

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.406}{\sqrt{2}} = \frac{3.406}{1.414} = 2.408$$

Για να δούμε αν θα βρούμε την ίδια τιμή εάν υπολογίσουμε το τυπικό σφάλμα ανατρέχοντας σε όλους τους μέσους όρους των 10 δειγμάτων μεγέθους 2

TABLE 7.2
Possible samples and sample means for
samples of size 2

Sample	Heights	$\bar{x}$
A, B	76, 78	77.0
A, C	76, 79	77.5
A, D	76, 81	78.5
A, E	76, 86	81.0
B, C	78, 79	78.5
B, D	78, 81	79.5
B, E	78, 86	82.0
C, D	79, 81	80.0
C, E	79, 86	82.5
D, E	81, 86	83.5

Υπενθυμίζουμε ότι ο μέσος όρος αυτού του πληθυσμού είναι

$$\mu = \mu_x = 80$$

Πάλι θα διαιρέσουμε με n και όχι με $(n-1)$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$$

Επειδή έχουμε αρκετά δεδομένα, ας υπολογίσουμε ξεχωριστά τον αριθμητή του κλάσματος που είναι μέσα στη ρίζα

$$\begin{aligned} & (77 - 80)^2 + (77.5 - 80)^2 + (78.5 - 80)^2 + (81 - 80)^2 + (78.5 - 80)^2 + \\ & (79.5 - 80)^2 + (82 - 80)^2 + (80 - 80)^2 + (82.5 - 80)^2 + (83.5 - 80)^2 \\ & = (-3)^2 + (-2.5)^2 + (-1.5)^2 + (1)^2 + (-1.5)^2 + \\ & \quad (-0.5)^2 + (2)^2 + (0)^2 + (2.5)^2 + (3.5)^2 \\ & = 9 + 6.25 + 2.25 + 1 + 2.25 + 0.25 + 4 + 0 + 6.25 + 12.25 \\ & = 43.5 \end{aligned}$$

Οπότε η τυπική απόκλιση είναι

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}} = \sqrt{\frac{43.5}{10}} = \sqrt{4.35} = 2.086$$

Βλέπουμε λοιπόν ότι η πραγματική τιμή της τυπικής απόκλισης της δειγματοληπτικής κατανομής του μέσου όρου των 10 δειγμάτων μεγέθους 2, δηλαδή το **2.086**, δεν είναι ίση με την τιμή που υπολογίσαμε με τον τύπο του τυπικού σφάλματος, **2.408**!

Ο λόγος που διαφέρουν είναι ότι ο τύπος του τυπικού σφάλματος που δώσαμε προηγουμένως

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

ισχύει μόνο όταν έχουμε επανατοποθέτηση (replacement), δηλαδή όταν μια μονάδα που επιλέγεται για ένα δείγμα, μπορεί να επιλεγεί ξανά για το ίδιο δείγμα!

- Δηλαδή για το προηγούμενο παράδειγμα, θα μπορούσε ένα δείγμα να είναι το (A, A), δηλαδή να εμφανιστεί δύο φορές ο ίδιος παίκτης σε ένα δείγμα

Αυτό δεν είναι ρεαλιστικό σε πολλές περιπτώσεις δειγματοληψίας, οπότε σε περίπτωση που δεν έχουμε επανατοποθέτηση, το τυπικό σφάλμα δίνεται από τον παρακάτω τύπο

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$$

όπου N είναι το μέγεθος του πληθυσμού και n είναι το μέγεθος του δείγματος. Αυτός ο διορθωμένος τύπος μας δίνει πράγματι τη σωστή τιμή του τυπικού σφάλματος της δειγματοληπτικής κατανομής του μέσου όρου για δείγματα μεγέθους n=2 (με τον πληθυσμό να έχει μέγεθος N=5)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{3.406}{\sqrt{2}} \cdot \sqrt{\frac{5-2}{5-1}} = \frac{3.406}{1.414} \cdot \sqrt{\frac{3}{4}} = \frac{3.406}{1.414} \cdot \sqrt{0.75} = 2.086$$

Βέβαια, όταν ο πληθυσμός είναι πολύ μεγαλύτερος από το δείγμα, οι δύο αυτοί τύποι του τυπικού σφάλματος δίνουν περίπου το ίδιο αποτέλεσμα

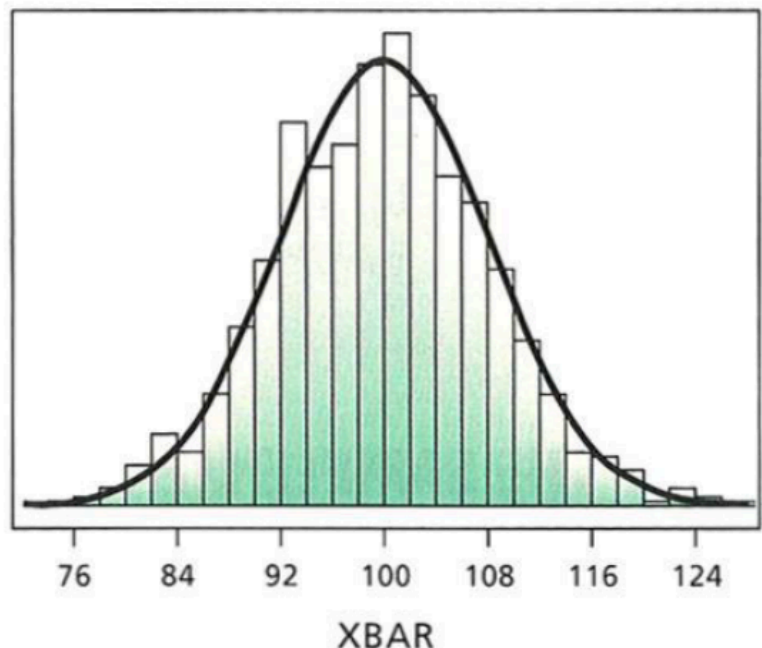
9.2.1. Κεντρικό οριακό θεώρημα

Όταν η μεταβλητή την οποία εξετάζουμε, π.χ. ύψος παικτών μπάσκετ, είναι κανονικά κατανεμημένη στον πληθυσμό, τότε και η δειγματοληπτική κατανομή του μέσου όρου είναι επίσης κανονικά κατανεμημένη

- Και μάλιστα, όταν μια μεταβλητή x είναι κανονικά κατανεμημένη με μέση τιμή μ και τυπική απόκλιση σ , τότε η δειγματοληπτική κατανομή του μέσου όρου της θα είναι επίσης κανονικά κατανεμημένη με μέση τιμή μ και τυπική απόκλιση $\sigma/\sqrt{n}$
 - Όπου n είναι το μέγεθος του δείγματος

PRINTOUT 7.1

Histogram of the sample means for 1000 samples of four IQs with superimposed normal curve

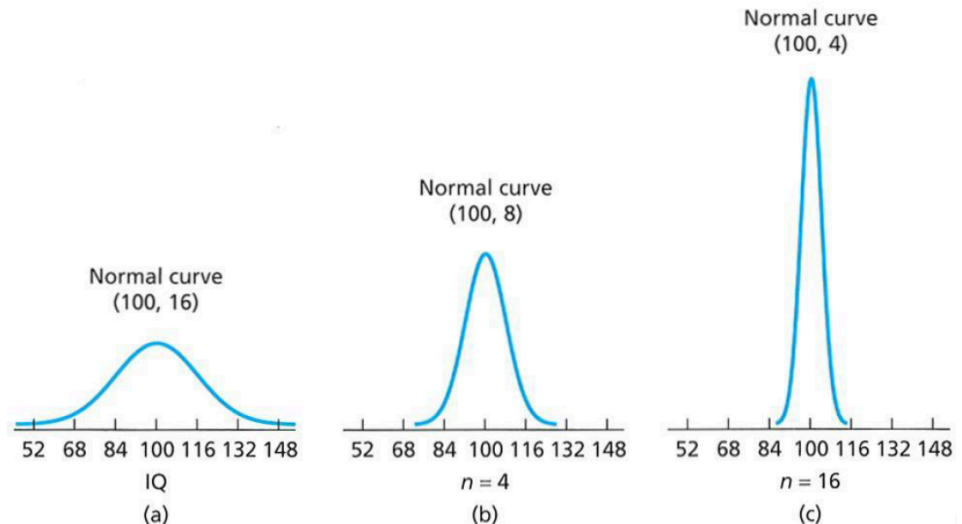


Όπως είδαμε στο προηγούμενο παράδειγμα, όσο μεγαλώνει το μέγεθος του δείγματος, τόσο μικραίνει η τυπική απόκλιση της δειγματοληπτικής κατανομής του μέσου όρου

- Δηλαδή οι μέσοι όροι των δειγμάτων βρίσκονται πιο κοντά στον μέσο όρο του πληθυσμού

FIGURE 7.4

(a) Normal distribution for IQs;
(b) sampling distribution of the sample mean for $n = 4$; (c) sampling distribution of the sample mean for $n = 16$



Τι γίνεται όμως όταν η μεταβλητή x δεν είναι κατανομημένη κανονικά στον πληθυσμό;

- Σύμφωνα λοιπόν με το **Κεντρικό Οριακό Θεώρημα** (Central Limit Theorem), η δειγματοληπτική κατανομή του μέσου όρου είναι κατά προσέγγιση κατανομημένη κανονικά ακόμα και όταν η μεταβλητή x ακολουθεί μια άλλη κατανομή!
 - Και μάλιστα η προσέγγιση αυτή γίνεται ακριβέστερη για μεγαλύτερη δείγματα

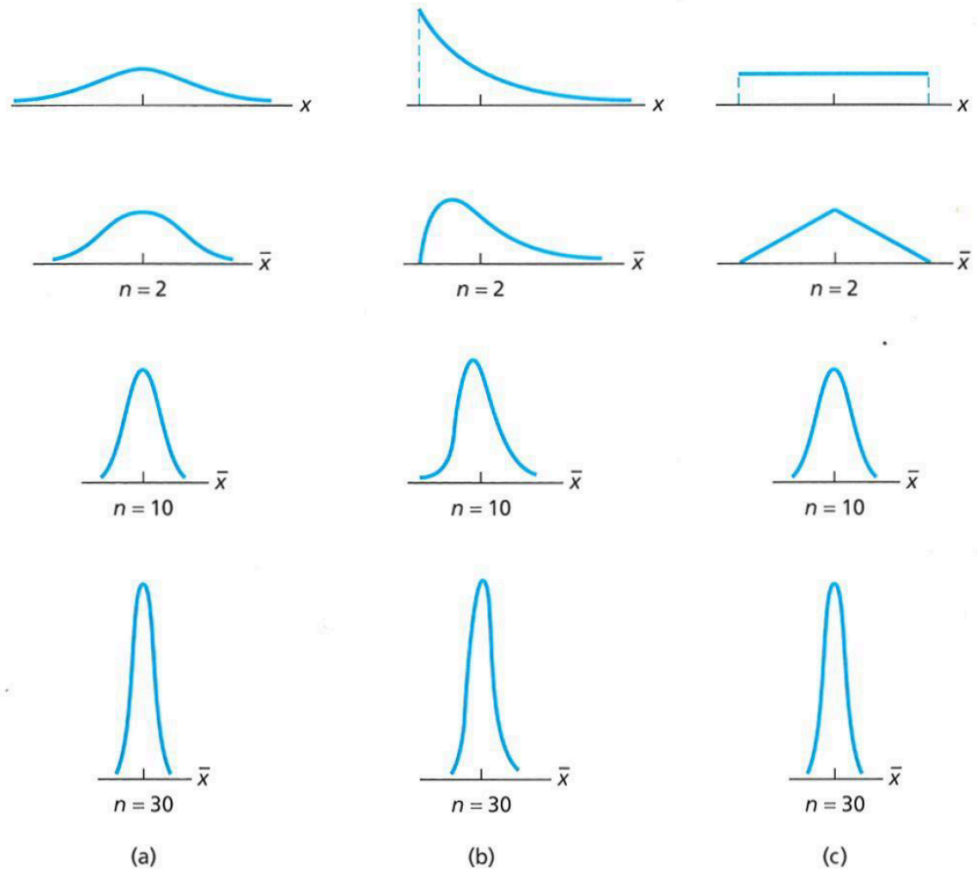
The Central Limit Theorem (CLT)

For a relatively large sample size, the variable $\bar{x}$ is approximately normally distributed, regardless of the distribution of the variable under consideration. The approximation becomes better and better with increasing sample size.

Κεντρικό οριακό θεώρημα (Central limit theorem)

FIGURE 7.6

Sampling distributions for (a) normal,
(b) reverse-J-shaped, and (c) uniform
variables



Ενδιαφέρουσες online προσομοιώσεις του Κεντρικού Οριακού Θεωρήματος

- <http://www.itcconline.net/green/java/Statistics/clt/cltsimulation.html>
- https://onlinestatbook.com/stat_sim/sampling_dist/

10. Διαστήματα εμπιστοσύνης μέσου όρου

Σημειακή εκτίμηση (point estimate)

- Συνοδεύεται από επίπεδο εμπιστοσύνης 0%

Point Estimate

A *point estimate* of a parameter is the value of a statistic used to estimate the parameter.

Διαστήματα εμπιστοσύνης (confidence intervals)

- Συνοδεύονται από κυμαινόμενο επίπεδο εμπιστοσύνης
 - 95% η πιο συνήθης τιμή (και αυτή που θα διαλέγετε εάν δεν ζητείται άλλο επίπεδο)
 - 99% και 90% επίσης σε χρήση

Confidence-Interval Estimate; Confidence Level

A *confidence-interval estimate* of a parameter consists of an interval of numbers obtained from a point estimate of the parameter and a percentage that specifies how confident we are that the parameter lies in the interval. The confidence percentage is called the *confidence level*.

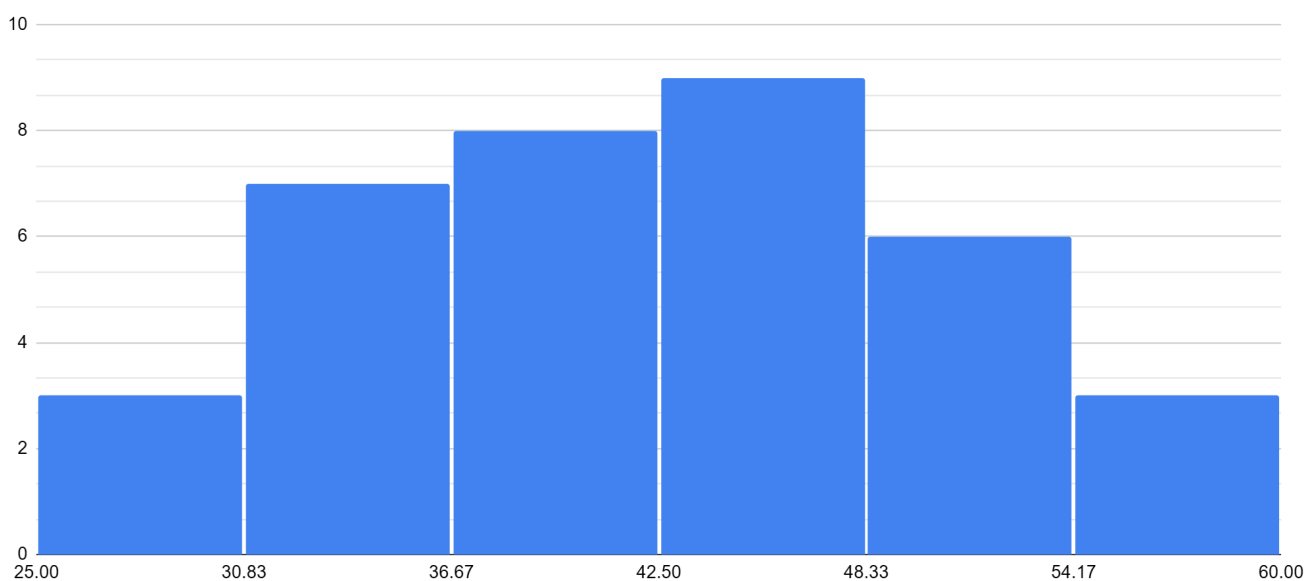
της μέσης τιμής

TABLE 8.1
Prices (\$1000s) of 36 randomly selected
new mobile homes

46.8	42.9	41.6	36.7	47.4	41.2	46.1	45.7	38.2
34.6	52.4	40.7	35.9	51.9	42.7	34.5	28.3	50.2
34.9	55.7	51.9	38.1	30.3	55.8	28.9	43.3	32.7
39.6	35.5	43.0	35.0	53.5	36.9	49.4	42.8	56.9

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1522.0}{36} = 42.28$$

$$s=7.83$$



Για να πάρουμε μια πρώτη ιδέα της έννοιας των διαστημάτων εμπιστοσύνης, εξετάσουμε τα διαστήματα εμπιστοσύνης που προκύπτουν από τον εμπειρικό κανόνα

(1) Διάστημα $(\mu - \sigma, \mu + \sigma) = (42.28 - 7.83, 42.28 + 7.83) = (34.45, 50.11)$ σε **επίπεδο εμπιστοσύνης 68.26%**

(2) Διάστημα $(\mu - 2\sigma, \mu + 2\sigma) = (42.28 - 2 \times 7.83, 42.28 + 2 \times 7.83) = (26.62, 57.94)$ σε **επίπεδο εμπιστοσύνης 95.44%**

(3) Διάστημα $(\mu - 3\sigma, \mu + 3\sigma) = (42.28 - 3 \times 7.83, 42.28 + 3 \times 7.83) = (18.79, 65.77)$ σε **επίπεδο εμπιστοσύνης 99.74%**

Τα διαστήματα εμπιστοσύνης της μέσης τιμής όμως, που θα μας απασχολήσουν σε αυτή την ενότητα είναι άλλη υπόθεση

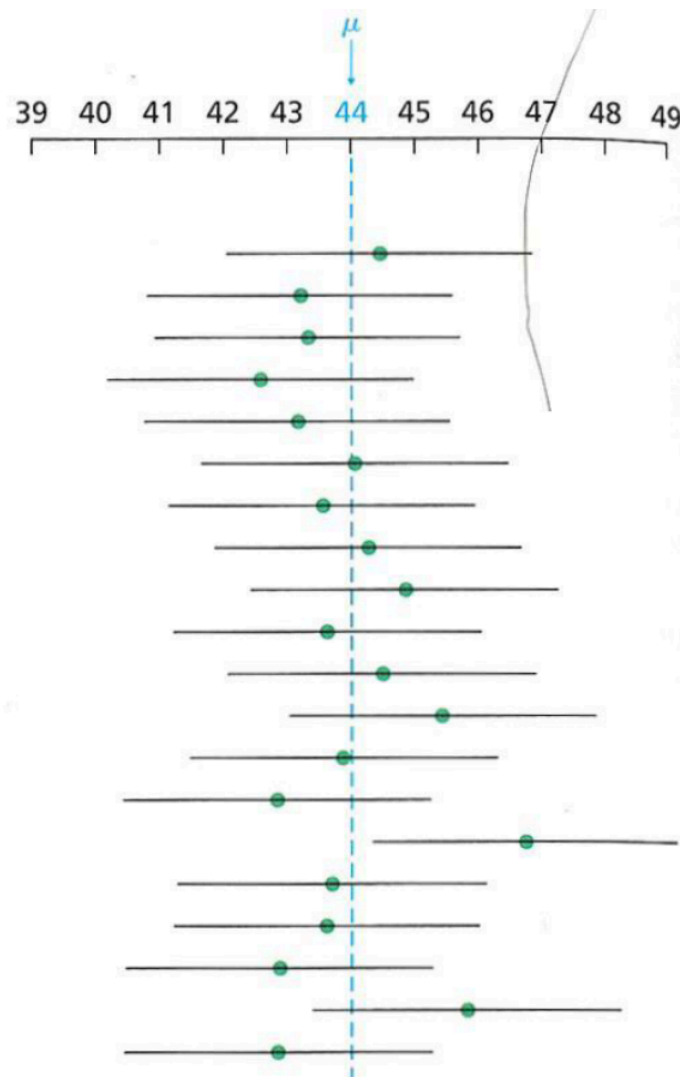
Ας δούμε τώρα τι συμβαίνει με διάφορα δείγματα και το μέσο όρο τους σε σχέση με τον (άγνωστο) μέσο όρο του πληθυσμού

FIGURE 8.1
Twenty confidence intervals for the mean price of all new mobile homes, each based on a sample of 36 new mobile homes

Τα διαστήματα εμπιστοσύνης είναι κεντραρισμένα στον μέσο όρο του δείγματος

- Άρα αλλάζουν από δείγμα σε δείγμα!

Sample	$\bar{x}$	95.44% CI	μ in CI?
1	44.45	42.06 to 46.85	yes
2	43.21	40.81 to 45.61	yes
3	43.33	40.93 to 45.73	yes
4	42.59	40.19 to 44.99	yes
5	43.17	40.77 to 45.57	yes
6	44.07	41.67 to 46.47	yes
7	43.56	41.16 to 45.96	yes
8	44.28	41.88 to 46.68	yes
9	44.87	42.48 to 47.27	yes
10	43.61	41.22 to 46.01	yes
11	44.51	42.11 to 46.91	yes
12	45.45	43.05 to 47.85	yes
13	43.88	41.48 to 46.28	yes
14	42.85	40.45 to 45.25	yes
15	46.73	44.33 to 49.13	no
16	43.70	41.30 to 46.10	yes
17	43.60	41.20 to 46.00	yes
18	42.88	40.48 to 45.28	yes
19	45.82	43.42 to 48.22	yes
20	42.84	40.45 to 45.24	yes



Ας δούμε τώρα την κανονική διαδικασία για υπολογισμό διαστημάτων εμπιστοσύνης

10.1. Διάστημα εμπιστοσύνης του μέσου όρου κατά την κατανομή z

Κατά την κατανομή z

The One-Sample z-Interval Procedure for a Population Mean

Assumptions

1. Normal population or large sample
2. σ known

Step 1 For a confidence level of $1 - \alpha$, use Table II to find $z_{\alpha/2}$.

Step 2 The confidence interval for μ is from

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{to} \quad \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is found in Step 1, n is the sample size, and $\bar{x}$ is computed from the sample data.

Step 3 Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

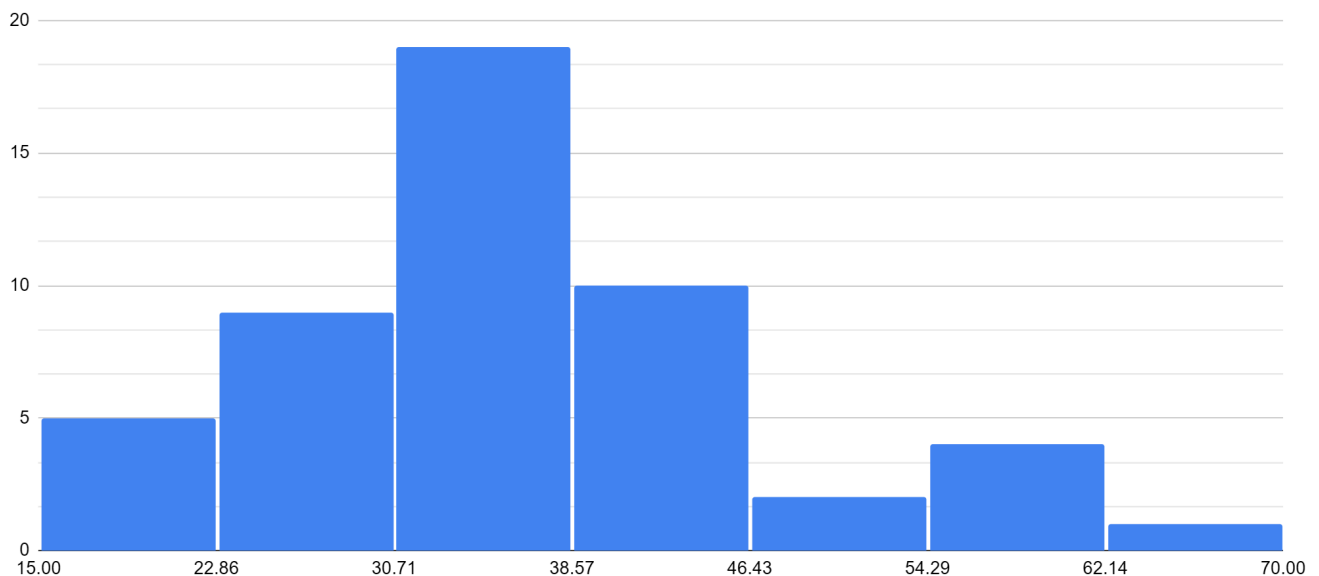
When to Use the z-Interval Procedure²

- For small samples—say, of size less than 15—the z-interval procedure should be used only when the variable under consideration is normally distributed or very close to being so.
- For samples of moderate size—say, between 15 and 30—the z-interval procedure can be used unless the data contain outliers or the variable under consideration is far from being normally distributed.
- For large samples—say, of size 30 or more—the z-interval procedure can be used essentially without restriction. However, if outliers are present and their removal is not justified, the effect of the outliers on the confidence interval should be examined; that is, you should compare the confidence intervals obtained with and without the outliers. If the effect is substantial, using a different procedure or taking another sample probably is best.
- If outliers are present but their removal is justified and results in a data set for which the z-interval procedure is appropriate (as previously stated), the procedure can be used.

TABLE 8.3

Ages of 50 randomly selected people in the
civilian labor force

22	58	40	42	43
32	34	45	38	19
33	16	49	29	30
43	37	19	21	62
60	41	28	35	37
51	37	65	57	26
27	31	33	24	34
28	39	43	26	38
42	40	31	34	38
35	29	33	32	33



Χαρακτηριστικά του δείγματος

$$\bar{x} = 36.38$$

$$s = 11.07$$

Επίσης είναι γνωστή η τυπική απόκλιση του πληθυσμού (κάτι πολύ ασυνήθιστο)

$$\sigma = 12.1$$

Step 1 For a confidence level of $1 - \alpha$, use Table II to find $z_{\alpha/2}$.

We want a 95% confidence interval, so $\alpha = 1 - 0.95 = 0.05$. From Table II, $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$.

Step 2 The confidence interval for μ is from

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{to} \quad \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

We have $\sigma = 12.1$, $n = 50$, and, from Step 1, $z_{\alpha/2} = 1.96$. To compute $\bar{x}$ for the data in Table 8.3, we apply the usual formula:

$$\bar{x} = \frac{\sum x}{n} = \frac{1819}{50} = 36.4,$$

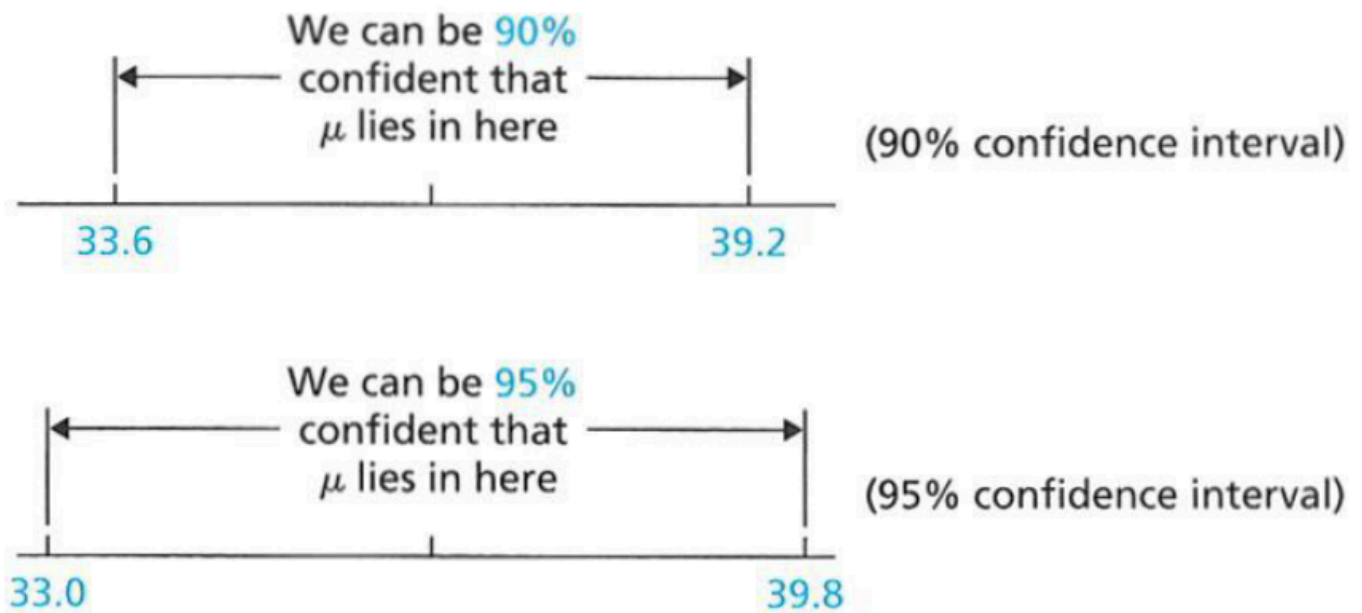
to one decimal place. Consequently, a 95% confidence interval for μ is from

$$36.4 - 1.96 \cdot \frac{12.1}{\sqrt{50}} \quad \text{to} \quad 36.4 + 1.96 \cdot \frac{12.1}{\sqrt{50}},$$

or 33.0 to 39.8.

Step 3 Interpret the confidence interval.

We can be 95% confident that the mean age, μ , of all people in the civilian labor force is somewhere between 33.0 years and 39.8 years.



Confidence and Precision

For a fixed sample size, decreasing the confidence level increases the precision, and vice-versa.

10.2. Διάστημα εμπιστοσύνης του μέσου όρου κατά την κατανομή t

Όταν η τυπική απόκλιση του πληθυσμού είναι άγνωστη

- Χρησιμοποιείται η [κατανομή t](#) (student ή t distribution) αντί της z (κανονικής)

FIGURE 8.7
Finding the t -value having area 0.05 to its right

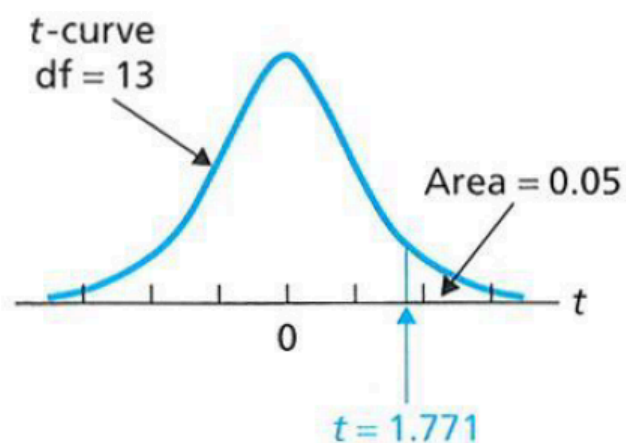
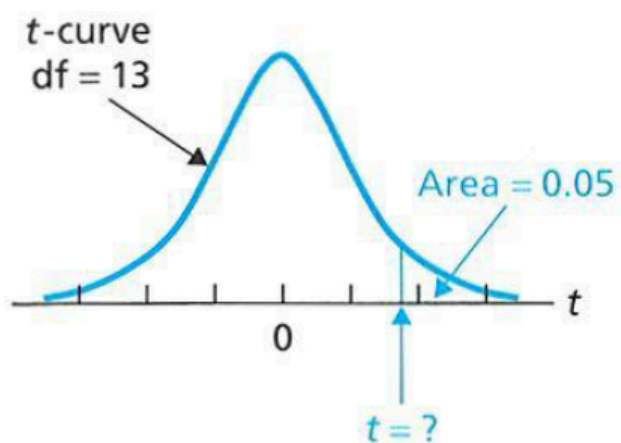


TABLE 8.4
Values of t_α

df	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$	df
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Η διαδικασία για τον υπολογισμό διαστήματος εμπιστοσύνης του μέσου όρου κατά την κατανομή t

The One-Sample t -Interval Procedure for a Population Mean

Assumptions

1. Normal population or large sample
2. σ unknown

Step 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $df = n - 1$, where n is the sample size.

Step 2 The confidence interval for μ is from

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}$ is found in Step 1 and $\bar{x}$ and s are computed from the sample data.

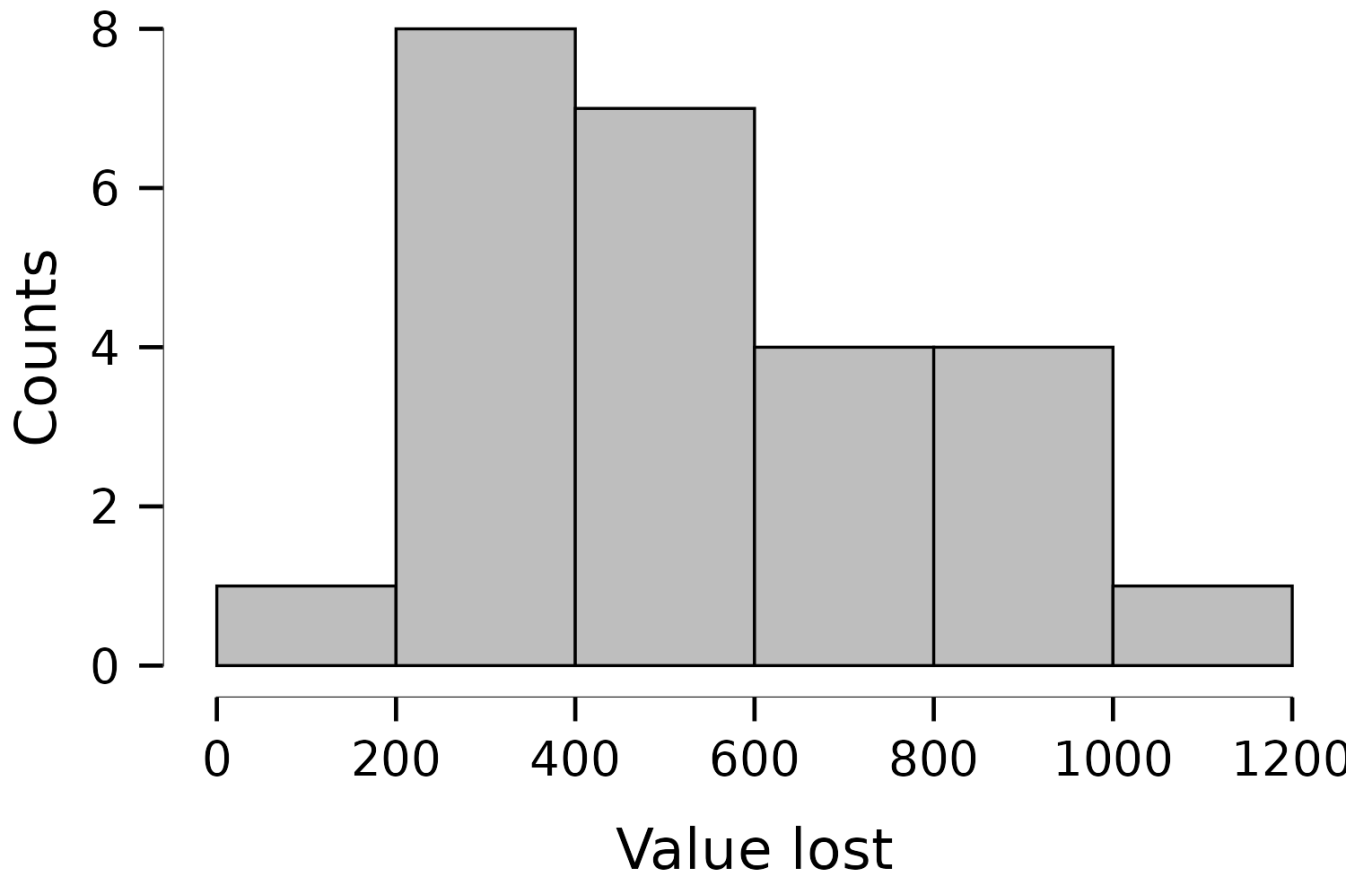
Step 3 Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

TABLE 8.5

Value lost (\$) for a sample of
25 pick-pocket offenses

447	207	627	430	883
313	844	253	397	214
217	768	1064	26	587
833	277	805	653	549
649	554	570	223	443



$n = 25$
 $\min = 26$
 $\max = 1064$
 $Q_1 = 277$
 $Q_3 = 653$
 $\text{median} = 549$
 $\bar{x} = 513.32$
 $s = 262.23$

Step 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n - 1$, where n is the sample size.

We want a 95% confidence interval, so $\alpha = 1 - 0.95 = 0.05$. For $n = 25$, $\text{df} = 25 - 1 = 24$. From Table IV, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.064$.

Step 2 The confidence interval for μ is from

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

From Step 1, $t_{\alpha/2} = 2.064$. Applying the usual formulas for $\bar{x}$ and s to the data in Table 8.5, we get $\bar{x} = 513.32$ and $s = 262.23$. Consequently, a 95% confidence interval for μ is from

$$513.32 - 2.064 \cdot \frac{262.23}{\sqrt{25}} \quad \text{to} \quad 513.32 + 2.064 \cdot \frac{262.23}{\sqrt{25}},$$

or 405.07 to 621.57.

(405.07, 621.57)

Παράδειγμα που δουλέψαμε αναλυτικά στην τάξη

TABLE 8.6

Sample of year's chicken consumption (lb)

47	39	62	49	50	70
59	53	55	0	65	63
53	51	50	72	45	

Επίδειξη βημάτων στην τάξη

1. OCR online
2. Convert in Word
3. Paste in DESCSTAT
4. Υπολογισμός βασικών μέτρων θέσεως και διασποράς
5. Έλεγχος ακραίων τιμών
 - ο Εσωτερικοί και εξωτερικοί φράχτες
 - ο Εμπειρικός κανόνας
6. Υπολογισμός διαστήματος εμπιστοσύνης

$$n = 17$$

$$\bar{x} = 51.94$$

$$s = 16.084$$

$$CL = 90\%$$

$$\alpha = 0.1$$

$$df = n-1 = 16$$

$$t_{a/2} = \pm 1.746$$

$$(45.129, 58.751)$$

11. Πίνακες αληθείας

μηδενική (null) και εναλλακτική (alternative) υπόθεση

Null hypothesis: A hypothesis to be tested. We use the symbol H_0 to represent the null hypothesis.

Alternative hypothesis: A hypothesis to be considered as an alternative to the null hypothesis. We use the symbol H_a to represent the alternative hypothesis.

H_0 : **The null hypothesis:** It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

$$H_0: \mu = \mu_0$$

H_a : **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0 .

Αμφίπλευρος έλεγχος (two-tailed test)

$$H_a: \mu \neq \mu_0$$

Μονόπλευρος έλεγχος (one-tailed test)

$$H_a: \mu < \mu_0$$

$$H_a: \mu > \mu_0$$

$H_0: \mu = 454$ g (the packaging machine is working properly)

$H_a: \mu \neq 454$ g (the packaging machine is not working properly)

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

H_0 : The drug reduces cholesterol by 25%. $p = 0.25$

H_a : The drug does not reduce cholesterol by 25%. $p \neq 0.25$

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

$H_0: \mu = 2.0$

$H_a: \mu \neq 2.0$

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol ($=$, $\neq$, $\geq$, $<$, $\leq$, $>$) for the null and alternative hypotheses. $H_0: \mu \text{ ___ } 66$ $H_a: \mu \text{ ___ } 66$

1. $H_0 : \mu = 66$

2. $H_a : \mu \neq 66$

We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

$H_0: \mu \geq 5$

$H_a: \mu < 5$

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol ($=$, $\neq$, $\geq$, $<$, $\leq$, $>$) for the null and alternative hypotheses.

$H_0: \mu \text{ ___ } 45$ $H_a: \mu \text{ ___ } 45$

1. $H_0: \mu \geq 45$

2. $H_a: \mu < 45$

In an issue of *U.S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

$$H_0: p \leq 0.066$$

$$H_a: p > 0.066$$

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol ($=$, $\neq$, $\geq$, $<$, $\leq$, $>$) for the null and alternative hypotheses.

$$H_0: p \text{ ___ } 0.40 \quad H_a: p \text{ ___ } 0.40$$

$$1. H_0: p = 0.40$$

$$2. H_a: p > 0.40$$

H_0	H_a
equal ($=$)	not equal ($\neq$) or greater than ($>$) or less than ($<$)
greater than or equal to ($\geq$)	less than ($<$)
less than or equal to ($\leq$)	more than ($>$)

σφάλματα τύπου I και II

πιθανότητα α και β

Decision:

H_0 is:		True	False
Decision:	Do not reject H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

• Truth tables



		H_0 T	H_0 F
	H_0 T	⊙ ✓ ⊙	II ~ b
	H_0 F	I ~ a	⊙ ✓ ⊙

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β



		DECISION	
		Reject H_0	Fail to Reject H_0
ACTUAL	H_0 True	Type I Error <i>Producer Risk</i> α -Risk False Positive	Correct Decision Confidence Interval = $1 - \alpha$
	H_a True	Correct Decision Power = $1 - \beta$	Type II Error <i>Consumer Risk</i> β -Risk False Negative

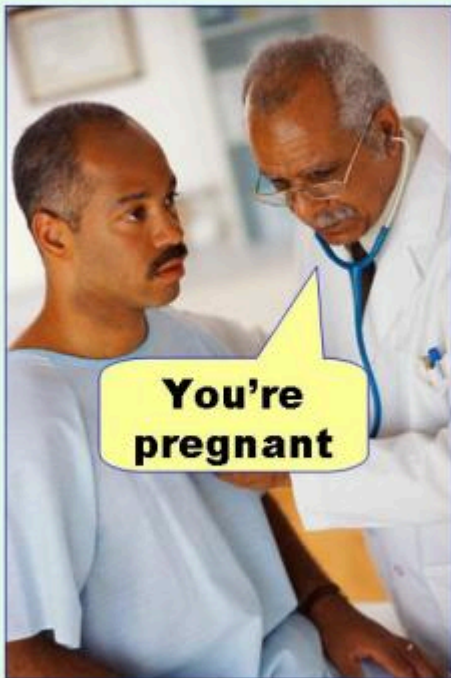
H_0 : Null Hypothesis H_a : Alternative Hypothesis

Type I and Type II Errors

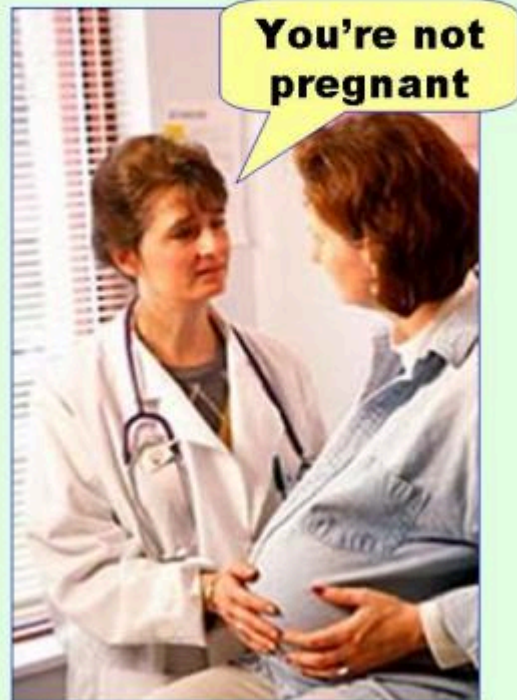
Type I error: Rejecting the null hypothesis when it is in fact true.

Type II error: Not rejecting the null hypothesis when it is in fact false.

Type I error
(false positive)



Type II error
(false negative)

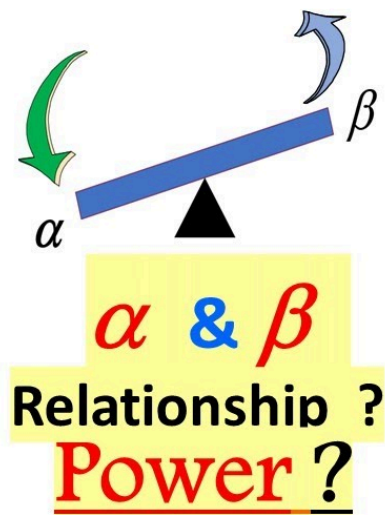


H_0 : you are not pregnant

H_a : you are pregnant!

Type 1 Error & Type 2 Error

VERDICT Set Free Jailed	 Type 1 Error (α)	 Type 2 Error (β)
	 Innocent	 Guilty
TRUTH		



Επίπεδο εμπιστοσύνης (confidence level) = $1 - \alpha$

Significance Level

The probability of making a Type I error, that is, of rejecting a true null hypothesis, is called the *significance level*, α , of a hypothesis test.

$$\text{Power} = 1 - P(\text{Type II error}) = 1 - \beta$$

Relation Between Type I and Type II Error Probabilities

For a fixed sample size, the smaller we specify the significance level, α , the larger will be the probability, β , of not rejecting a false null hypothesis.

Possible Conclusions for a Hypothesis Test

Suppose that a hypothesis test is conducted at a small significance level.

- If the null hypothesis is rejected, we conclude that the alternative hypothesis is true.
- If the null hypothesis is not rejected, we conclude that the data do not provide sufficient evidence to support the alternative hypothesis.

H_0 : Defendant is innocent

H_a : Defendant is guilty.

- a. Explain the meaning of a Type I error.
- b. Explain the meaning of a Type II error.
- c. If you were the defendant, would you want α to be large or small? Explain your answer.
- d. If you were the prosecuting attorney, would you want β to be large or small? Explain your answer.
- e. What are the consequences to the court system if you make $\alpha = 0$? $\beta = 0$?

12. Έλεγχοι υποθέσεων

Υπάρχουν τα ακόλουθα κύρια είδη ελέγχων υποθέσεων

1. Έλεγχοι μέσης τιμής

- Θα εξετάσουμε ελέγχους μέσης τιμής ενός πληθυσμού και ελέγχους που συγκρίνουν τις μέσες τιμές δύο πληθυσμών

2. Έλεγχοι διασποράς

- Δεν θα ασχοληθούμε καθόλου με τέτοιους ελέγχους

3. Έλεγχο ποσοστών

- Θα εξετάσουμε ελέγχους του ποσοστού ενός πληθυσμού

4. Έλεγχο προσαρμογής

- Θα εξετάσουμε μόνο τον έλεγχο χ^2

Βασική λογική όλων των ελέγχων υποθέσεων

1. Παίρνουμε ένα τυχαίο δείγμα από τον πληθυσμό.
2. Εάν τα δεδομένα του δείγματος είναι συμβατά με τη μηδενική υπόθεση, δεν απορρίπτουμε τη μηδενική υπόθεση
 - Πρέπει βέβαια να υπάρχει ένα σαφές κριτήριο για την απόρριψη της μηδενικής υπόθεσης
3. Εάν τα δεδομένα του δείγματος δεν είναι συμβατά με τη μηδενική υπόθεση, απορρίπτουμε τη μηδενική υπόθεση υπέρ της εναλλακτικής

Σημαντικές έννοιες

- Επίπεδο εμπιστοσύνης (confidence level), που συνήθως εκφράζεται ως ποσοστό τοις εκατό, π.χ. 95%
- Σημαντικότητα (α), που ισούται με 100 μείον το επίπεδο εμπιστοσύνης, και συνήθως εκφράζεται ως μέρος της μονάδας, π.χ. 0.05 (για επίπεδο εμπιστοσύνης 95%)

Τρεις τρόποι να διενεργήσουμε τον έλεγχο z

1. Με κρίσιμες τιμές (critical values)
2. Με υπολογισμό του p-value
3. Με διαστήματα εμπιστοσύνης

12.1. Έλεγχος z της μέσης τιμής

Έλεγχος z της μέσης τιμής (z-test)

Χρήση πινάκων κανονικής κατανομής

One-Mean z-Test

Purpose To perform a hypothesis test for a population mean, μ

Assumptions

1. Simple random sample
2. Normal population or large sample
3. σ known

Step 1 The null hypothesis is $H_0: \mu = \mu_0$, and the alternative hypothesis is

$$\begin{array}{ccc} H_a: \mu \neq \mu_0 & \text{or} & H_a: \mu < \mu_0 & \text{or} & H_a: \mu > \mu_0 \\ \text{(Two tailed)} & & \text{(Left tailed)} & & \text{(Right tailed)} \end{array}$$

Step 2 Decide on the significance level, α .

Step 3 Compute the value of the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

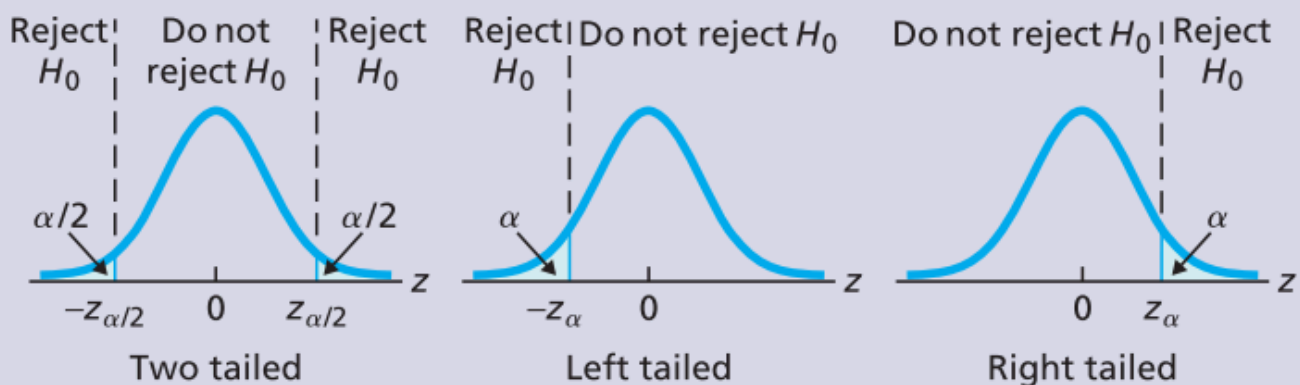
and denote that value z_0 .

CRITICAL-VALUE APPROACH

Step 4 The critical value(s) are

$\pm z_{\alpha/2}$ (Two tailed) or $-z_{\alpha}$ (Left tailed) or z_{α} (Right tailed)

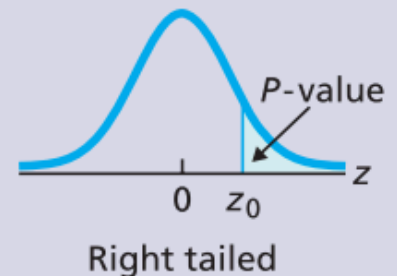
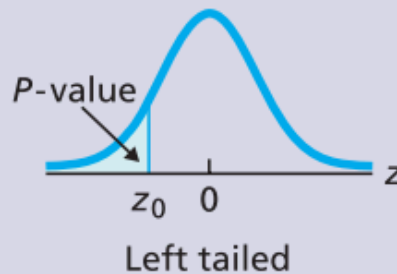
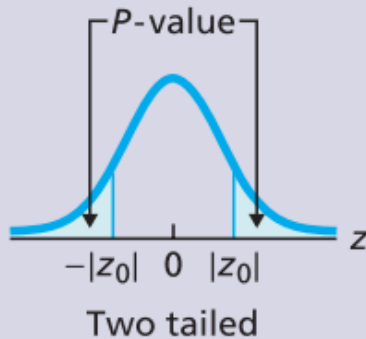
Use Table II to find the critical value(s).



Step 5 If the value of the test statistic falls in the rejection region, reject H_0 ; otherwise, do not reject H_0 .

P-VALUE APPROACH

Step 4 Use Table II to obtain the P -value.



Step 5 If $P \leq \alpha$, reject H_0 ; otherwise, do not reject H_0 .

Some important values of z_α

$z_{0.10}$	$z_{0.05}$	$z_{0.025}$	$z_{0.01}$	$z_{0.005}$
1.28	1.645	1.96	2.33	2.575

TABLE 9.8

Guidelines for using the P -value
to assess the evidence
against the null hypothesis

P -value	Evidence against H_0
$P > 0.10$	Weak or none
$0.05 < P \leq 0.10$	Moderate
$0.01 < P \leq 0.05$	Strong
$P \leq 0.01$	Very strong

Αμφίπλευρος έλεγχος

TABLE 9.11

Top speeds, in miles per hour,
for a sample of 35 cheetahs

57.3	57.5	59.0	56.5	61.3
57.6	59.2	65.0	60.1	59.7
62.6	52.6	60.7	62.3	65.2
54.8	55.4	55.5	57.8	58.7
57.8	60.9	75.3	60.6	58.1
55.9	61.6	59.6	59.8	63.4
54.7	60.2	52.4	58.3	66.0



$H_0: \mu = 60$ mph (mean top speed of cheetahs is 60 mph)

$H_a: \mu \neq 60$ mph (mean top speed of cheetahs is not 60 mph)

Τυπική απόκλιση της ταχύτητας όλων των cheetah $\sigma = 3.2$ mph

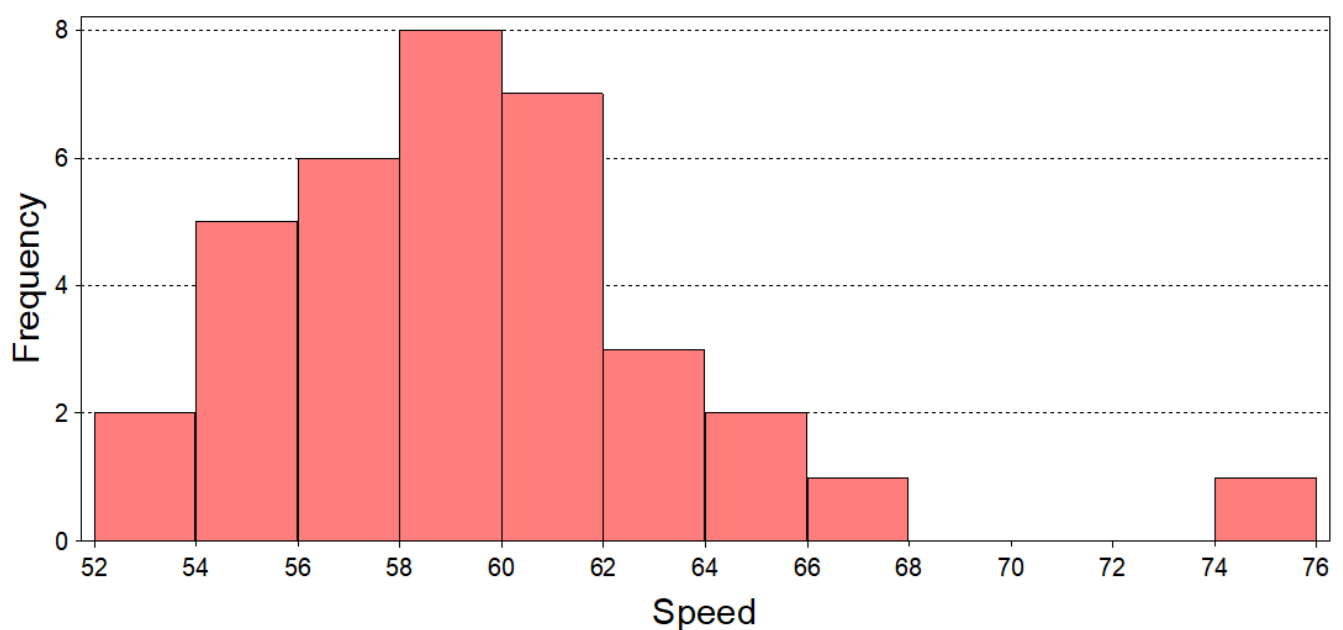
CL = 95%, $\alpha = 0.05$

CASE	Speed (mph)
1	52.4
2	52.6
3	54.7
4	54.8
5	55.4
6	55.5
7	55.9
8	56.5
9	57.3
10	57.5
11	57.6
12	57.8
13	57.8
14	58.1
15	58.3
16	58.7

17	59
18	59.2
19	59.6
20	59.7
21	59.8
22	60.1
23	60.2
24	60.6
25	60.7
26	60.9
27	61.3
28	61.6
29	62.3
30	62.6
31	63.4
32	65
33	65.2
34	66
35	75.3

Λόγω μεγέθους δείγματος, δεν είναι ανάγκη να κάνουμε έλεγχο κανονικότητας

- Ας ρίξουμε όμως μια ματιά στο ιστόγραμμα του δείγματος



Φαίνεται να υπάρχει ένα πολύ γρήγορο cheetah

DESCRIPTIVE STATISTICS		
n	35	large sample if $n \geq 30$
Min	52.4	
Max	75.3	
Range	22.9	Max-Min
Q_1	57.400	
Q_3	61.100	
Interquartile range (IQR)	3.700	$Q_3 - Q_1$
Median	59.200	
Mode	57.800	
Sum	2083.4	
Mean	59.526	Sum/n
Standard deviation (sample)	4.274	$\sqrt{\text{Variance calculated with } n-1}$
Variance (sample)	18.266	Calculated with $n-1$
CV (sample, %)	–	Standard deviation/Mean (%)
Standard deviation (population)	4.212	$\sqrt{\text{Variance calculated with } n}$
Variance (population)	17.744	Calculated with n
CV (population, %)	7.08	Standard deviation/Mean (%)
Skewness	1.401	0 for normal distribution

$H_0: \mu = 60$ mph (mean top speed of cheetahs is 60 mph)

$H_a: \mu \neq 60$ mph (mean top speed of cheetahs is not 60 mph)

$$\bar{x} = 59.526$$

$$\mu_0 = 60$$

$$\sigma = 3.2$$

$$n = 35$$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

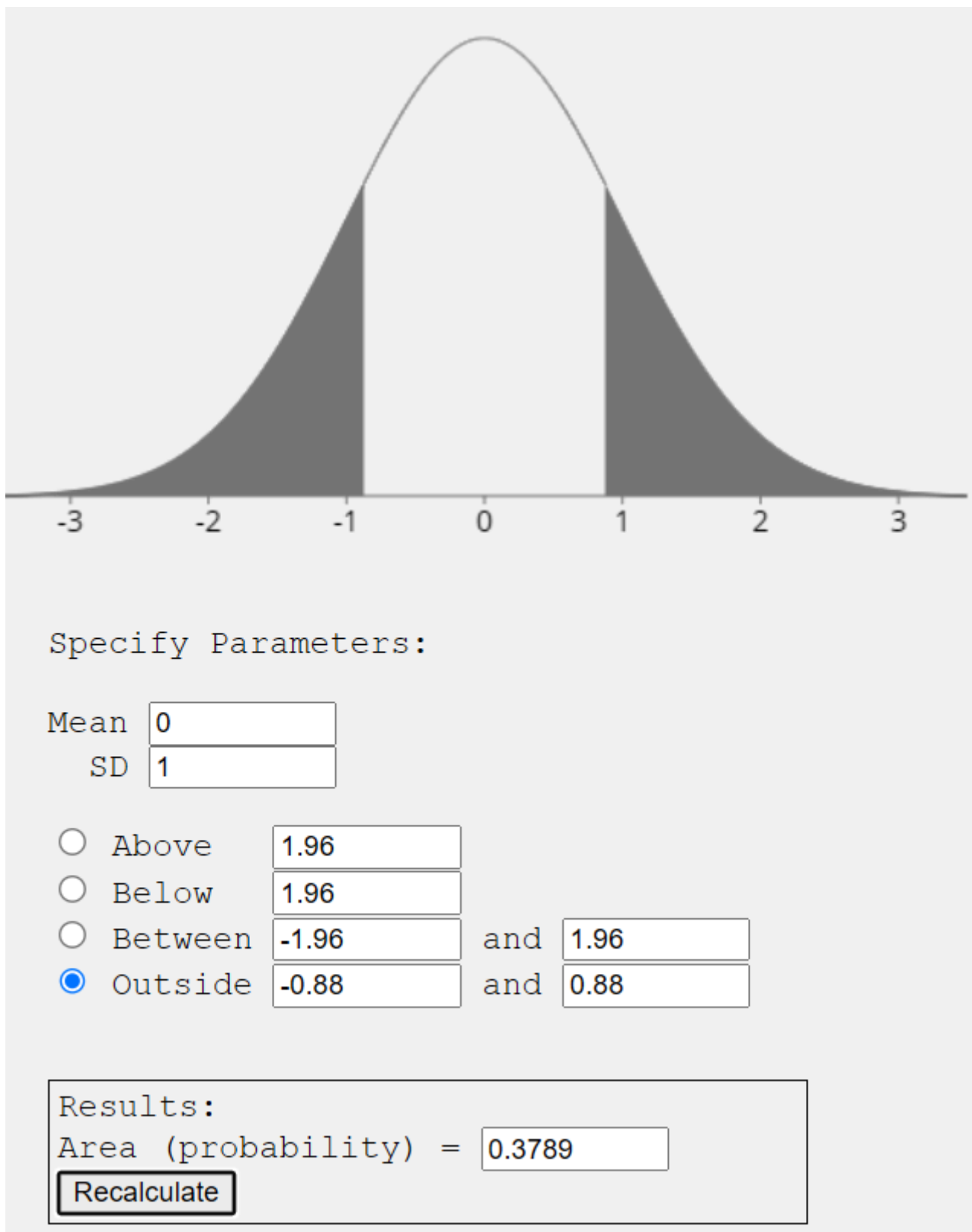
$$z = \frac{59.526 - 60}{3.2 / \sqrt{35}} = -0.88$$

Z TEST FOR THE MEAN		
Confidence level (%)	95	

z value left (two tail)	-1.96	
z value right (two tail)	1.96	
z value left (one tail)	-1.64	
z value right (one tail)	1.64	
Mean	59.526	
Standard deviation	4.274	
Population mean (μ_0)	60	
Population standard deviation (σ)	3.2	
z statistic	-0.88	$(\mu - \mu_0)/(\sigma/\sqrt{n})$
P[z < z statistic]	0.1903	from standard normal distribution
p value (two tail)	0.3806	
p value (one tail)	0.1903	

Το παρακάτω γράφημα από τη σελίδα

https://onlinestatbook.com/2/calculators/normal_dist.html



Μονόπλευρος (αριστερόστροφος) έλεγχος

Ο Jack, παίκτης golf, ισχυρίζεται ότι μπορεί να ρίξει τη μπάλα 275 yards (one yard = 91.44 cm)

Οι τελευταίες 25 επιδόσεις του εικονίζονται παρακάτω

Distances (yards) of 25 drives by Jack

266	254	248	249	297
261	293	261	266	279
222	212	282	281	265
240	284	253	274	243
272	279	261	273	295

Η τυπική απόκλιση όλων των επιδόσεων του παίκτη είναι $\sigma = 20$ yards
Δύσπιστος φίλος του Jack ζητάει τον παρακάτω έλεγχο

$H_0: \mu = 275$ yards (Jack's claim)

$H_a: \mu < 275$ yards (Jean's suspicion)

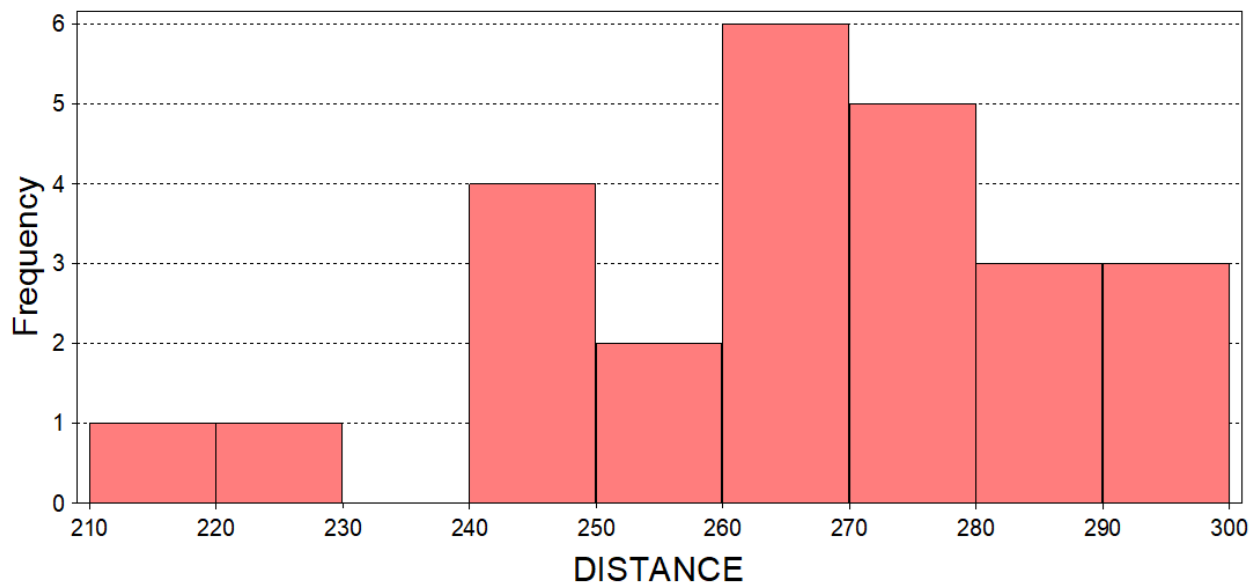
Ο έλεγχος είναι μονόπλευρος και μάλιστα αριστερόστροφος

Τα δεδομένα ταξινομούνται από το μικρότερο στο μεγαλύτερο ως εξής

CASE	DATA
1	212
2	222
3	240
4	243
5	248
6	249
7	253
8	254
9	261
10	261
11	261
12	265
13	266
14	266
15	272
16	273

17	274
18	279
19	279
20	281
21	282
22	284
23	293
24	295
25	297

Να επαληθεύσουμε ότι τα δεδομένα είναι κατανεμημένα κανονικά
Το παρακάτω ιστόγραμμα στο στατιστικό πρόγραμμα Statistix



Και με τον εμπειρικό κανόνα (DESCSTAT)

EMPIRICAL RULE		
Mean	264.400	
Standard deviation	21.231	
Mean - 3×Standard deviation	200.707	
Mean - 2×Standard deviation	221.938	
Mean - 1×Standard deviation	243.169	
Mean + 1×Standard deviation	285.631	
Mean + 2×Standard deviation	306.862	
Mean + 3×Standard deviation	328.093	

Number of data in (Mean $\pm$ 1×Standard deviation)	21	Out of 25 data points
Number of data in (Mean $\pm$ 2×Standard deviation)	24	Out of 25 data points
Number of data in (Mean $\pm$ 3×Standard deviation)	25	Out of 25 data points
% of data in (Mean $\pm$ 1×Standard deviation)	84.0	68.27% for normal distribution
% of data in (Mean $\pm$ 2×Standard deviation)	96.0	95.45% for normal distribution
% of data in (Mean $\pm$ 3×Standard deviation)	100.0	99.73% for normal distribution

Βασικά στατιστικά μεγέθη από το DESCSTAT

DESCRIPTIVE STATISTICS		
n	25	large sample if $n \geq 30$
Min	212	
Max	297	
Range	85	Max-Min
Q ₁	253.000	
Q ₃	279.000	
Interquartile range (IQR)	26.000	Q ₃ -Q ₁
Median	266.000	
Mode	261.000	
Sum	6610	
Mean	264.400	Sum/n
Standard deviation (sample)	21.231	√Variance calculated with n-1
Variance (sample)	450.750	Calculated with n-1
CV (sample, %)	–	Standard deviation/Mean (%)
Standard deviation (population)	20.802	√Variance calculated with n
Variance (population)	432.720	Calculated with n
CV (population, %)	7.87	Standard deviation/Mean (%)
Skewness	-0.676	0 for normal distribution

$$\bar{x} = 264.4$$

$$\mu_0 = 275$$

$$\sigma = 20$$

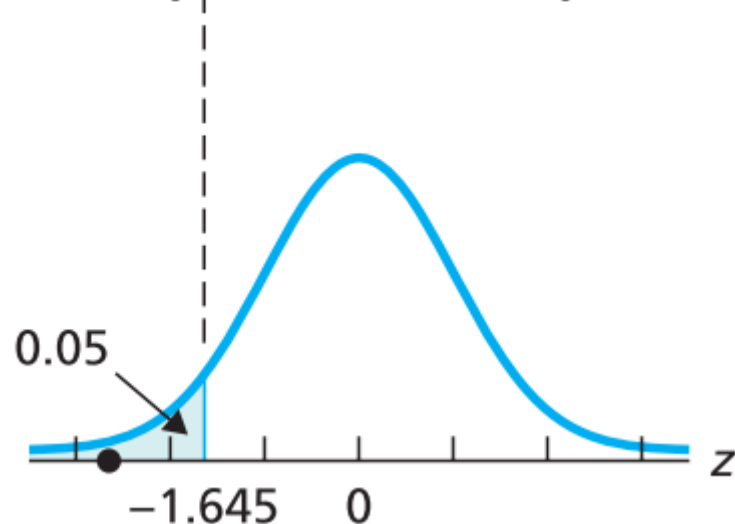
$$n = 25$$

$$\mu_{\bar{x}} = \mu = 275 \text{ yards} \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = 4 \text{ yards}$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 275}{4}$$

$$z = \frac{\bar{x} - 275}{4} = \frac{264.4 - 275}{4} = -2.65$$

Reject H_0 | Do not reject H_0



Διενέργεια του στατιστικού ελέγχου με το DESCSTAT

Z TEST FOR THE MEAN		
Confidence level (%)	95	
z value left (two tail)	-1.96	
z value right (two tail)	1.96	
z value left (one tail)	-1.64	
z value right (one tail)	1.64	
Mean	264.400	
Standard deviation	21.231	
Population mean (μ_0)	275	
Population standard deviation (σ)	20	
z statistic	-2.65	$(\mu - \mu_0)/(\sigma/\sqrt{n})$
$P[z < z \text{ statistic}]$	0.0040	from standard normal distribution
p value (two tail)	0.0080	
p value (one tail)	0.0040	

Διενέργεια του στατιστικού ελέγχου με το στατιστικό πρόγραμμα Minitab

Descriptive Statistics

N	Mean	StDev	SE Mean	95% Upper Bound for μ
---	------	-------	---------	---------------------------

25	264.40	21.23	4.00	270.98
----	--------	-------	------	--------

Known standard deviation = 20

Test

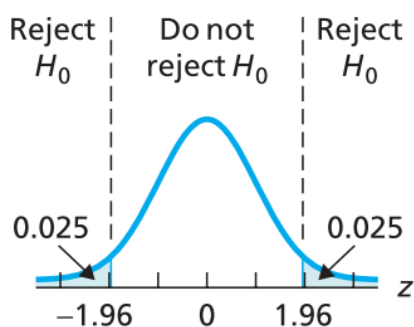
Null hypothesis $H_0: \mu = 275$

Alternative hypothesis $H_1: \mu < 275$

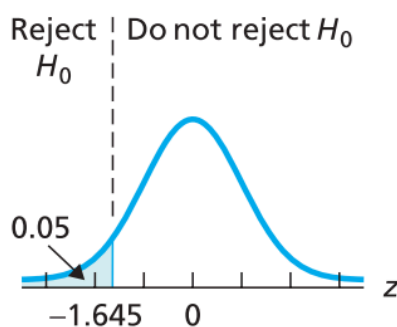
Z-Value	P-Value
---------	---------

-2.65	0.004
-------	-------

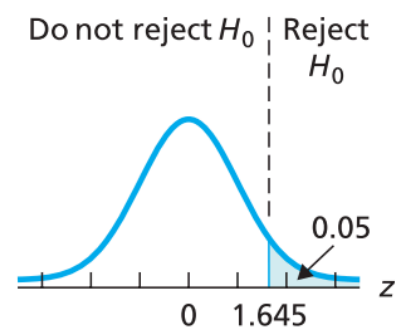
Μην ξεχνάμε τις κρίσιμες τιμές της κανονικής κατανομής για επίπεδο εμπιστοσύνης 95%



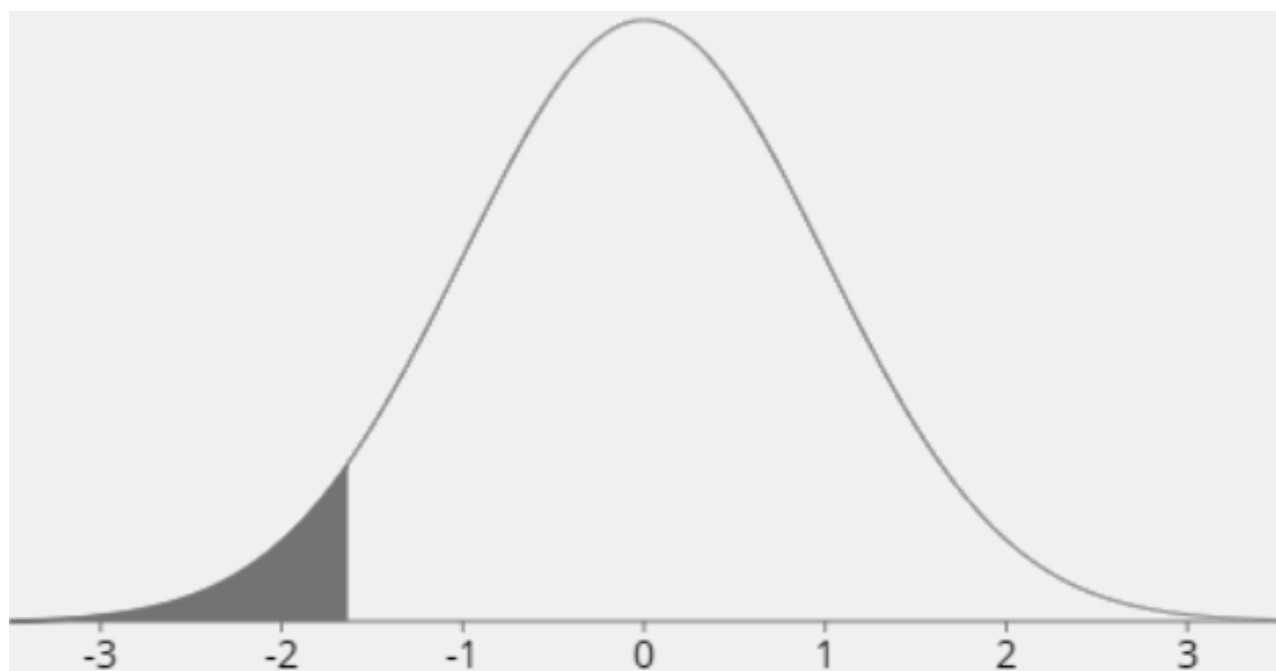
(a) Two tailed



(b) Left tailed



(c) Right tailed



Specify Parameters:

Mean
SD

- ☐ Above
☒ Below
☐ Between and
☐ Outside and

Results:

Area (probability) =

Μονόπλευρος (δεξιόστροφος) έλεγχος

TABLE 9.9

This year's prices, in dollars,
for 40 history books

82.55	72.80	73.89	80.54
80.26	74.43	81.37	82.28
77.55	88.25	73.58	89.23
74.35	77.44	78.91	77.50
77.83	77.49	87.25	98.93
74.25	82.71	78.88	78.25
80.35	77.45	90.29	79.42
67.63	91.48	83.99	80.64
101.92	83.03	95.59	69.26
80.31	98.72	87.81	69.20

CL = 99%, επομένως $\alpha = 0.01$

Τυπική απόκλιση πληθυσμού = 7.61

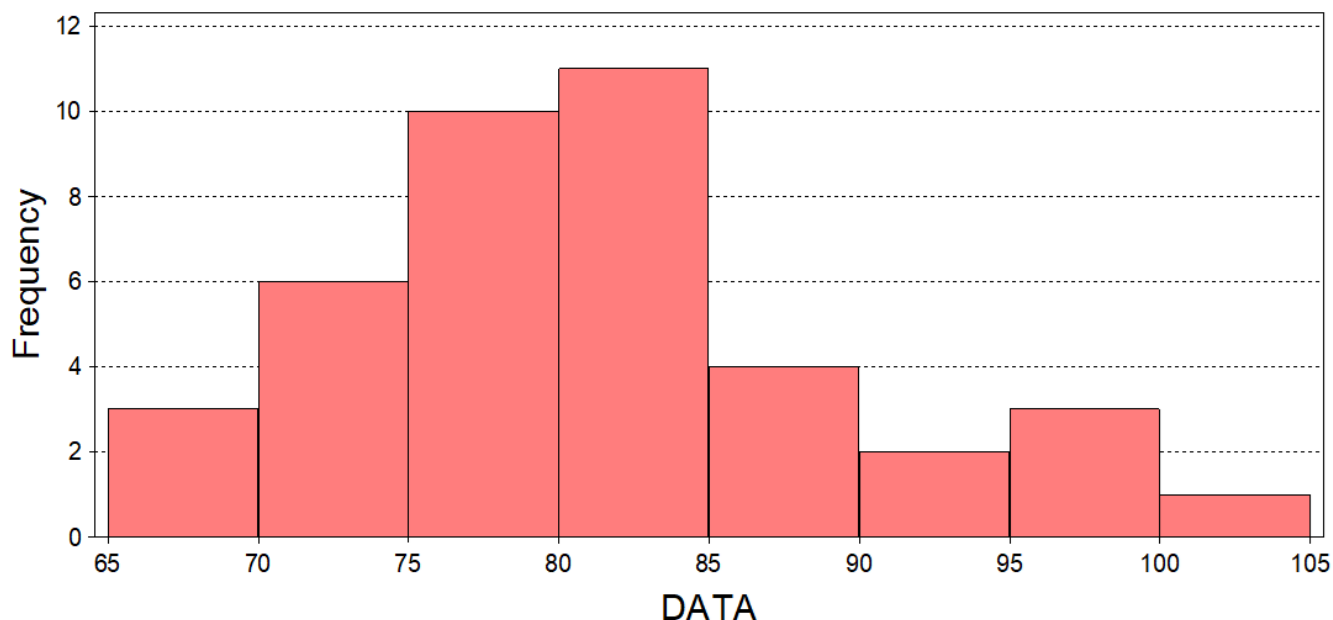
$H_0: \mu = \$78.01$ (mean price has not increased)

$H_a: \mu > \$78.01$ (mean price has increased).

Κανονικότητα δείγματος

- Αν και δεν χρειάζεται να ελέγξουμε γιατί το δείγμα είναι μεγάλο
 - Που σημαίνει ότι η δειγματοληπτική κατανομή του μέσου όρου θα είναι κανονική ακόμα και εάν ο πληθυσμός των δεδομένων δεν είναι

Πράγματι, όπως περιμένουμε για οικονομικά μεγέθη, το δείγμα των τιμών των βιβλίων εμφανίζει θετική στρέβλωση



Βασικά στατιστικά μεγέθη

DESCRIPTIVE STATISTICS			
n	40	large sample if $n \geq 30$	
Min	67.63		
Max	101.92		
Range	34.29	Max-Min	
Q_1	77.448		
Q_3	84.805		
Interquartile range (IQR)	7.357	$Q_3 - Q_1$	
Median	80.285		
Mode	none		
Sum	3257.61		
Mean	81.440	Sum/n	
Standard deviation (sample)	8.093	$\sqrt{\text{Variance calculated with } n-1}$	
Variance (sample)	65.498	Calculated with $n-1$	
CV (sample, %)	–	Standard deviation/Mean (%)	
Standard deviation (population)	7.991	$\sqrt{\text{Variance calculated with } n}$	
Variance (population)	63.860	Calculated with n	
CV (population, %)	9.81	Standard deviation/Mean (%)	
Skewness	0.792	0 for normal distribution	

Έλεγχος z

Z TEST FOR THE MEAN

Confidence level (%)	99	
z value left (two tail)	-2.58	
z value right (two tail)	2.58	
z value left (one tail)	-2.33	
z value right (one tail)	2.33	
Mean	81.440	
Standard deviation	8.093	
Population mean (μ_0)	78.01	
Population standard deviation (σ)	7.61	
z statistic	2.85	$(\mu - \mu_0)/(\sigma/\sqrt{n})$
P[z < z statistic]	0.9978	from standard normal distribution
p value (two tail)	0.0044	
p value (one tail)	0.0022	

$$\bar{x} = 81.44$$

$$\mu_0 = 78.01$$

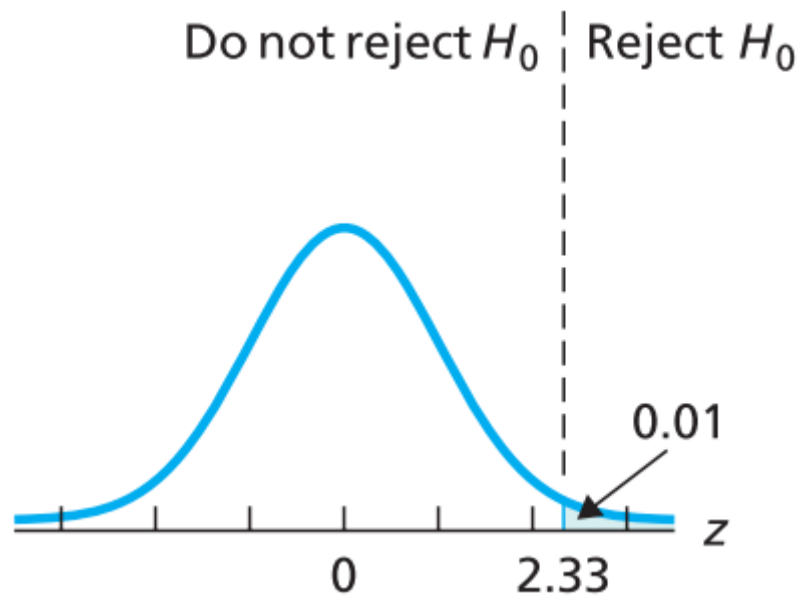
$$\sigma = 7.61$$

$$n = 40$$

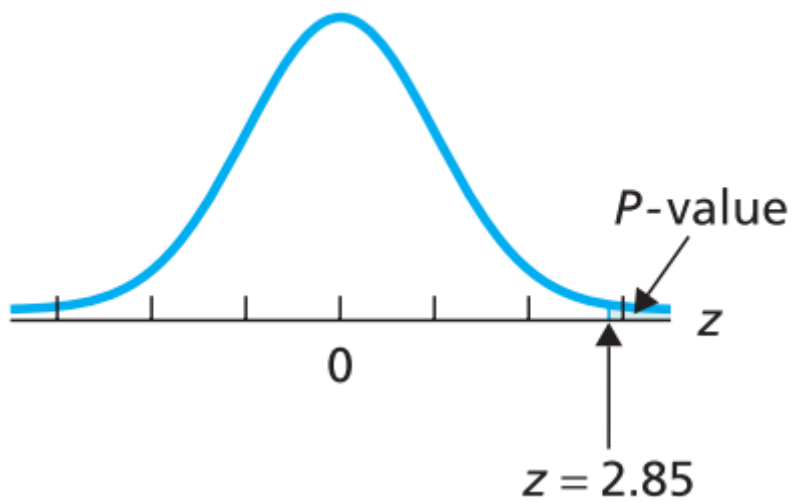
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

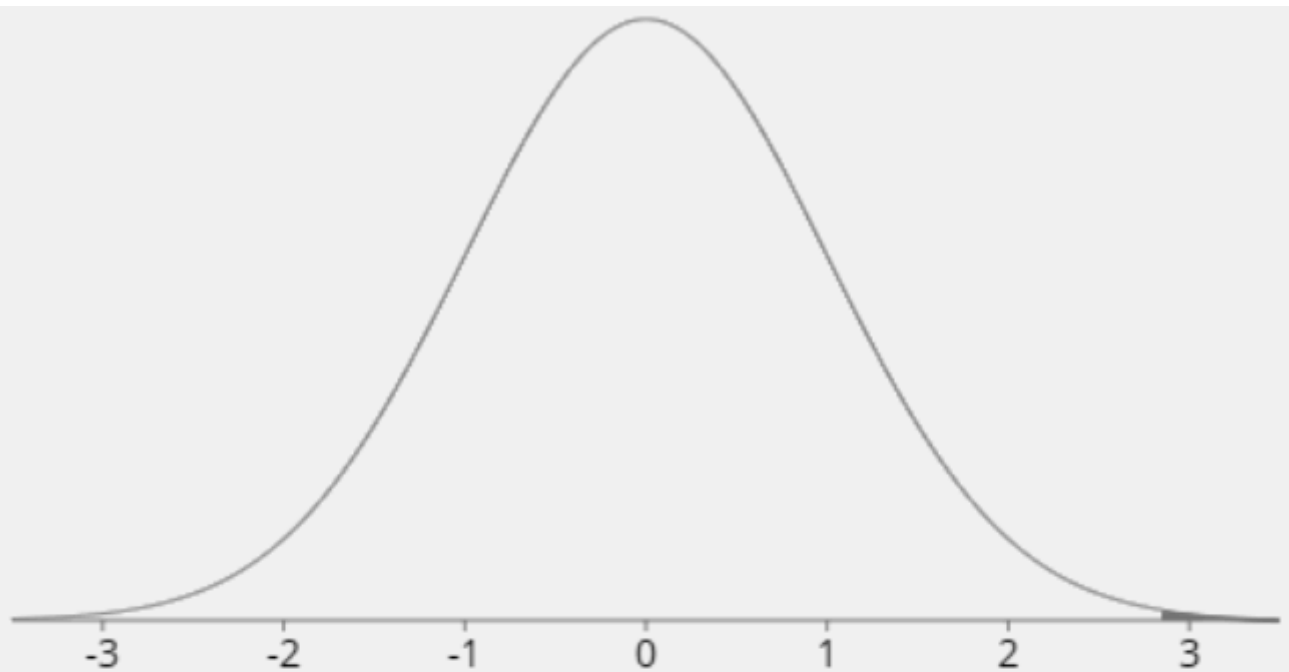
$$z = \frac{81.440 - 78.01}{7.61/\sqrt{40}} = 2.85$$

Με τη μέθοδο των κρίσιμων τιμών



Με τη μέθοδο της πιθανότητας p





Specify Parameters:

Mean
SD

- ☒ Above
☐ Below
☐ Between and
☐ Outside and

Results:

Area (probability) =

12.2. Έλεγχος t της μέσης τιμής

12.2.1. Ένα δείγμα

της μέσης τιμής

t-test

One-Mean t-Test

Purpose To perform a hypothesis test for a population mean, μ

Assumptions

1. Simple random sample
2. Normal population or large sample
3. σ unknown

Step 1 The null hypothesis is $H_0: \mu = \mu_0$, and the alternative hypothesis is

$$\begin{array}{ccc} H_a: \mu \neq \mu_0 & \text{or} & H_a: \mu < \mu_0 & \text{or} & H_a: \mu > \mu_0 \\ \text{(Two tailed)} & & \text{(Left tailed)} & & \text{(Right tailed)} \end{array}$$

Step 2 Decide on the significance level, α .

Step 3 Compute the value of the test statistic

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

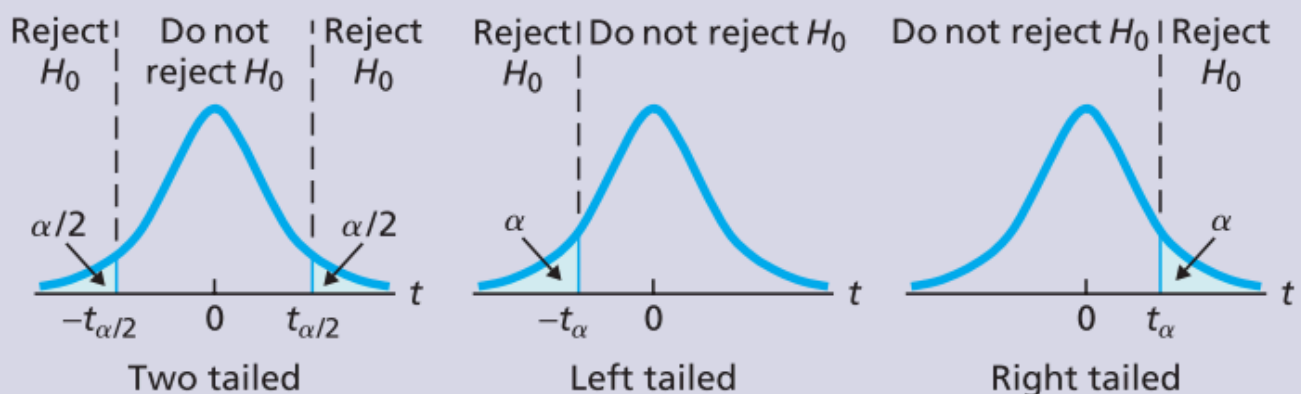
and denote that value t_0 .

CRITICAL-VALUE APPROACH

Step 4 The critical value(s) are

$\pm t_{\alpha/2}$ (Two tailed) or $-t_{\alpha}$ (Left tailed) or t_{α} (Right tailed)

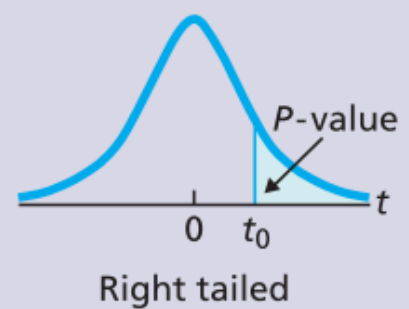
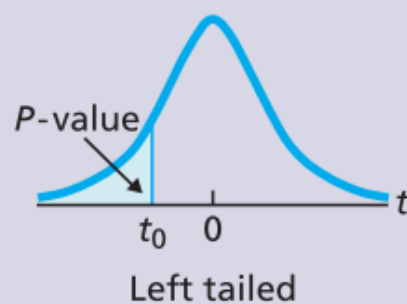
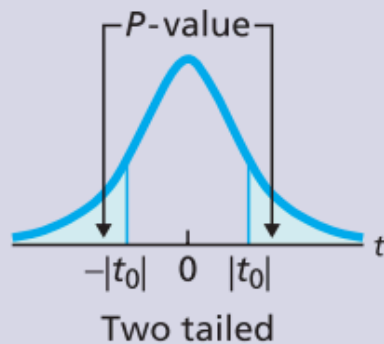
with $df = n - 1$. Use Table IV to find the critical value(s).



Step 5 If the value of the test statistic falls in the rejection region, reject H_0 ; otherwise, do not reject H_0 .

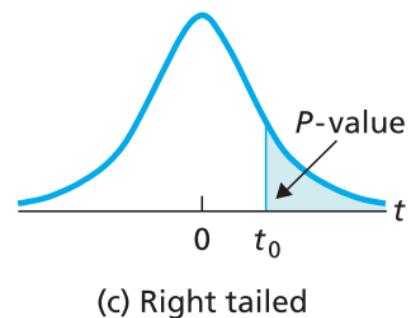
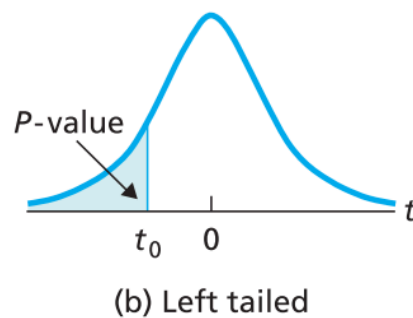
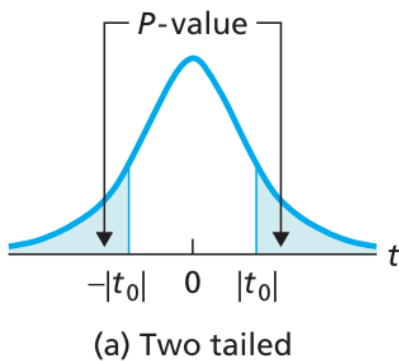
P-VALUE APPROACH

Step 4 The t -statistic has $df = n - 1$. Use Table IV to estimate the P -value, or obtain it exactly by using technology.



Step 5 If $P \leq \alpha$, reject H_0 ; otherwise, do not reject H_0 .

χρήση πινάκων κατανομής t

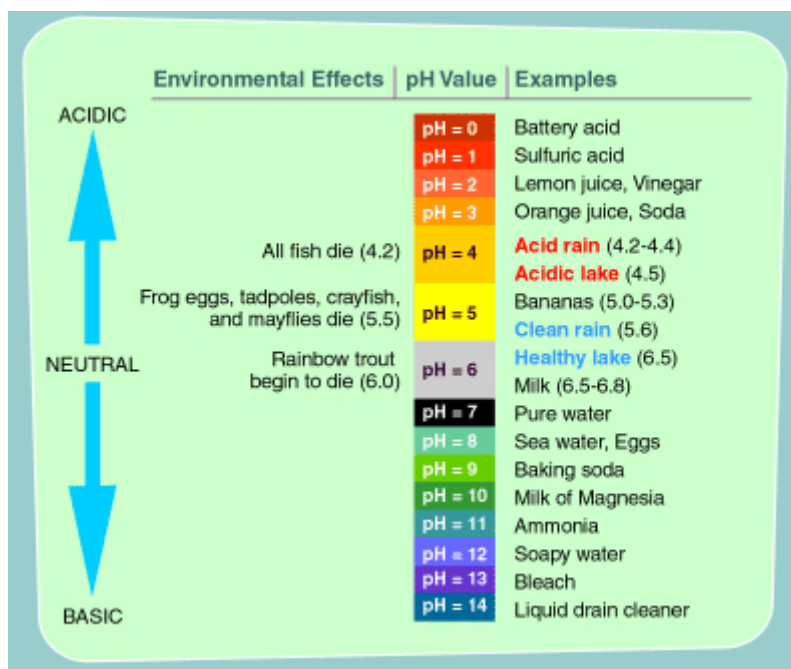
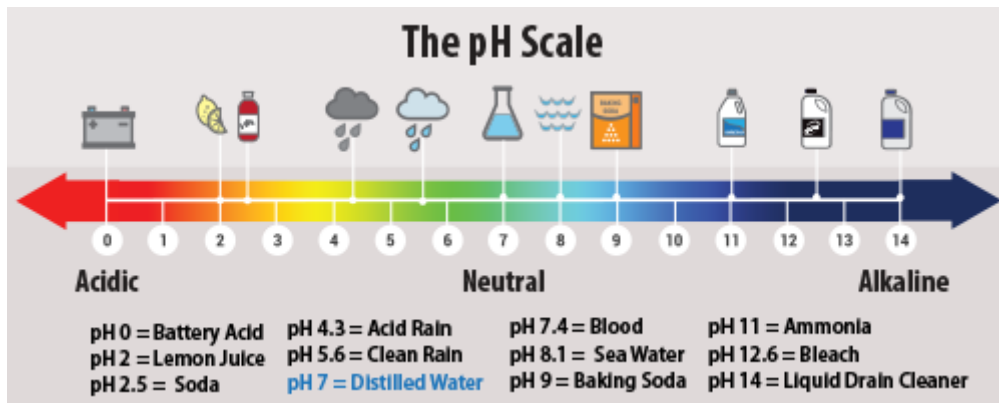
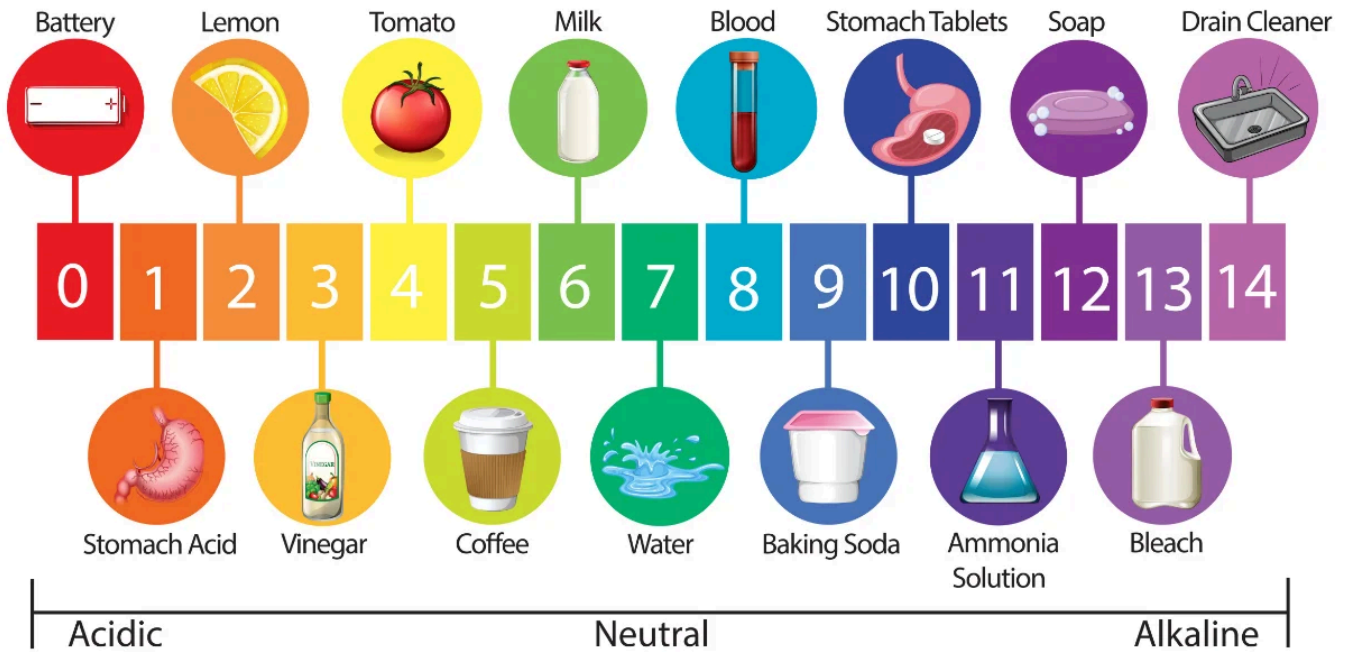


pH (ενεργός οξύτητα) λιμνών

pH levels for 15 lakes

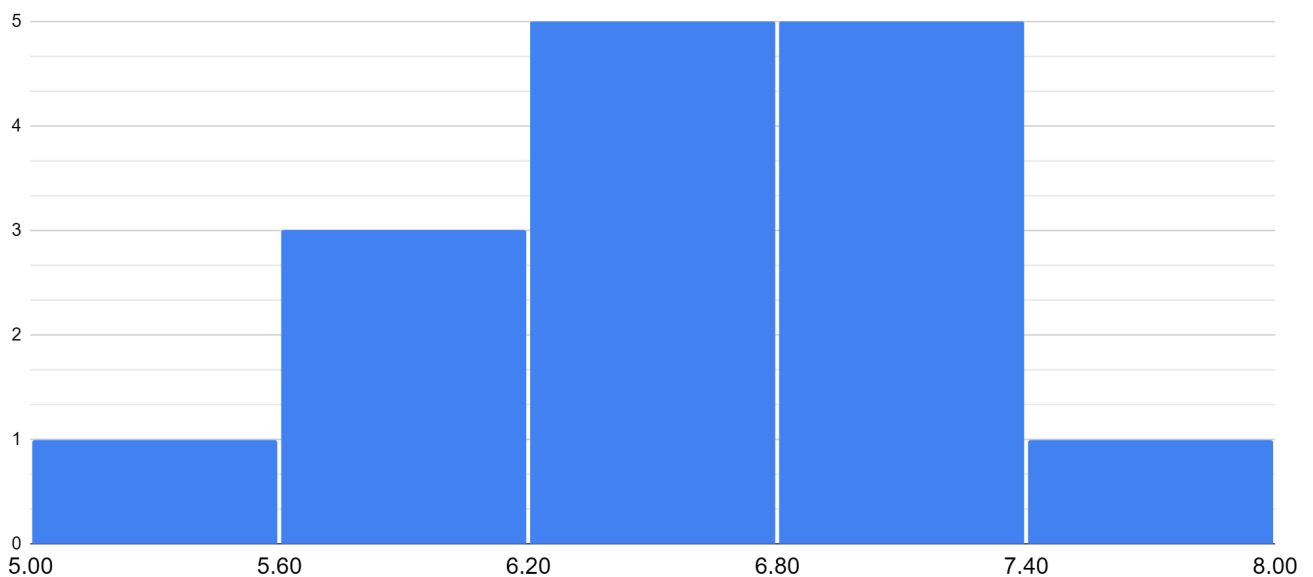
7.2	7.3	6.1	6.9	6.6
7.3	6.3	5.5	6.3	6.5
5.7	6.9	6.7	7.9	5.8

The pH Scale



1	5.5
2	5.7

3	5.8
4	6.1
5	6.3
6	6.3
7	6.5
8	6.6
9	6.7
10	6.9
11	6.9
12	7.2
13	7.3
14	7.3
15	7.9



Αντιγράφω τα παραπάνω δεδομένα στο [DESCSTAT](#) και βρίσκουμε

$$n = 15$$

$$\text{Min} = 5.5$$

$$\text{Max} = 7.9$$

$$Q_1 = 6.2 \text{ (δεν πειράζει που δεν είναι ίδιο με αυτό που βρίσκουμε με το χέρι)}$$

$$\text{Median} = 6.6$$

$$Q_3 = 7.05 \text{ (δεν πειράζει που δεν είναι ίδιο με αυτό που βρίσκουμε με το χέρι)}$$

$$\text{IQR} = 0.85$$

$$\text{Lower outer fence} = 3.65$$

$$\text{Lower inner fence} = 4.925$$

$$\text{Upper inner fence} = 8.325$$

Upper outer fence = 9.6

Mode = 6.3

(Sum = 99)

Mean = 6.6

Variance (sample) = 0.451

Standard deviation (sample) = 0.672

Για τον έλεγχο μέσου όρου είτε με την κατανομή z (αν ξέρουμε την τυπική απόκλιση του πληθυσμού) ή με την κατανομή t, παίρνουμε τις παρακάτω τιμές (από το DESCSTAT ή πίνακες)

Confidence level (%) = 95

z value lower (two tail) = 1.960

z value upper (two tail) = 1.960

z value left (one tail) = -1.645

z value right (one tail) = 1.645

Degrees of freedom = 14

t value lower (two tail) = -2.145

t value upper (two tail) = 2.145

t value left (one tail) = -1.761

t value right (one tail) = 1.761

t = -2.31, p = 0.037

12.2.2. Εξαρτημένα δείγματα (dependent samples ή paired t-test)

Standing and running pulse (revisited)

Τα δεδομένα που συλλέξαμε στην τάξη νωρίτερα στο εξάμηνο

CASE	NAME	SPULSE	RPULSE	DIFF	SMOKING	MALE
1	Αφροδίτη	54	94	40	0	0
2	Βασίλης	76	100	24	1	1
3	Βικτωρία	78	110	32	0	0
4	Γιώργος	100	110	10	1	1
5	Δανάη	102	156	54	0	0
6	Δημήτρης	56	96	40	0	1
7	Ευσεβία	84	140	56	0	0

8	Κωνσταντίνος	90	92	2	0	1
9	Μαρίζα	84	108	24	0	0

Περιγραφικά στατιστικά μεγέθη της μεταβλητής **SPULSE**

$$n = 9$$

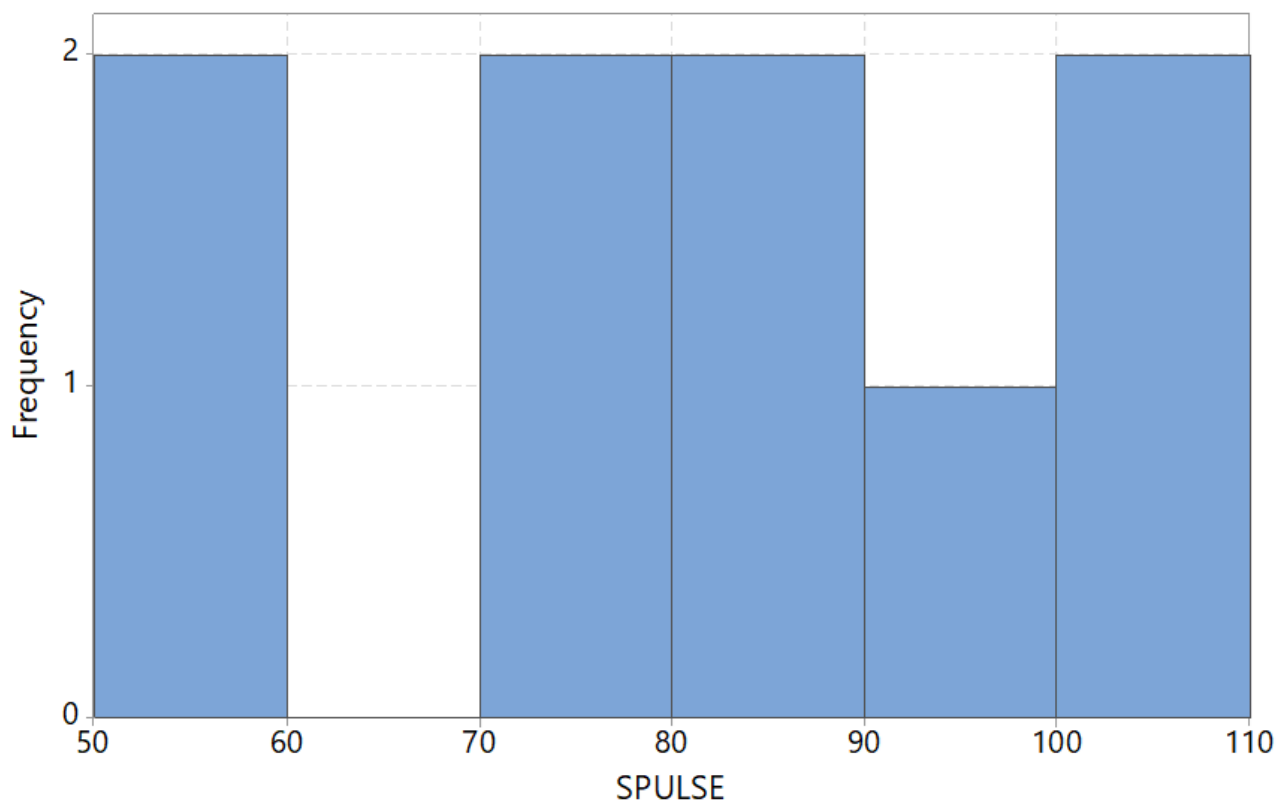
$$\text{Min} = 54$$

$$\text{Max} = 102$$

$$\sum_{i=1}^9 x_i = 724 \text{ και } \bar{x} = 80.444$$

$$s = 16.905$$

Το δείγμα είναι μικρό ($n = 9 < 30$), επομένως πρέπει να εξετάσουμε το ιστόγραμμα του



Περιγραφικά στατιστικά μεγέθη της μεταβλητής **RPULSE**

$$n = 9$$

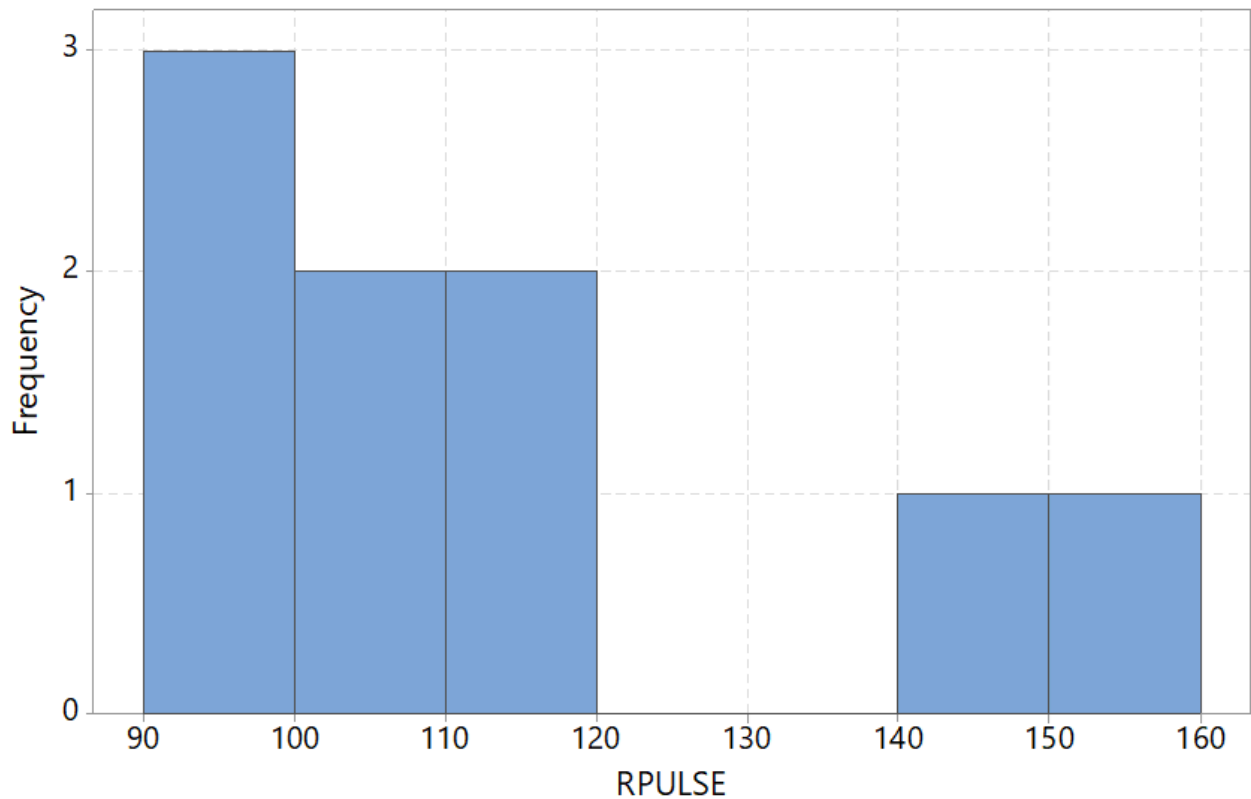
$$\text{Min} = 92$$

$$\text{Max} = 156$$

$$\sum_{i=1}^9 x_i = 1006 \text{ και } \bar{x} = 111.778$$

$$s = 21.987$$

Ιστόγραμμα



Τώρα θα δούμε το δείγμα των διαφορών **DIFF** (RPULSE–SPULSE)

Περιγραφικά στατιστικά μεγέθη

$$n = 9$$

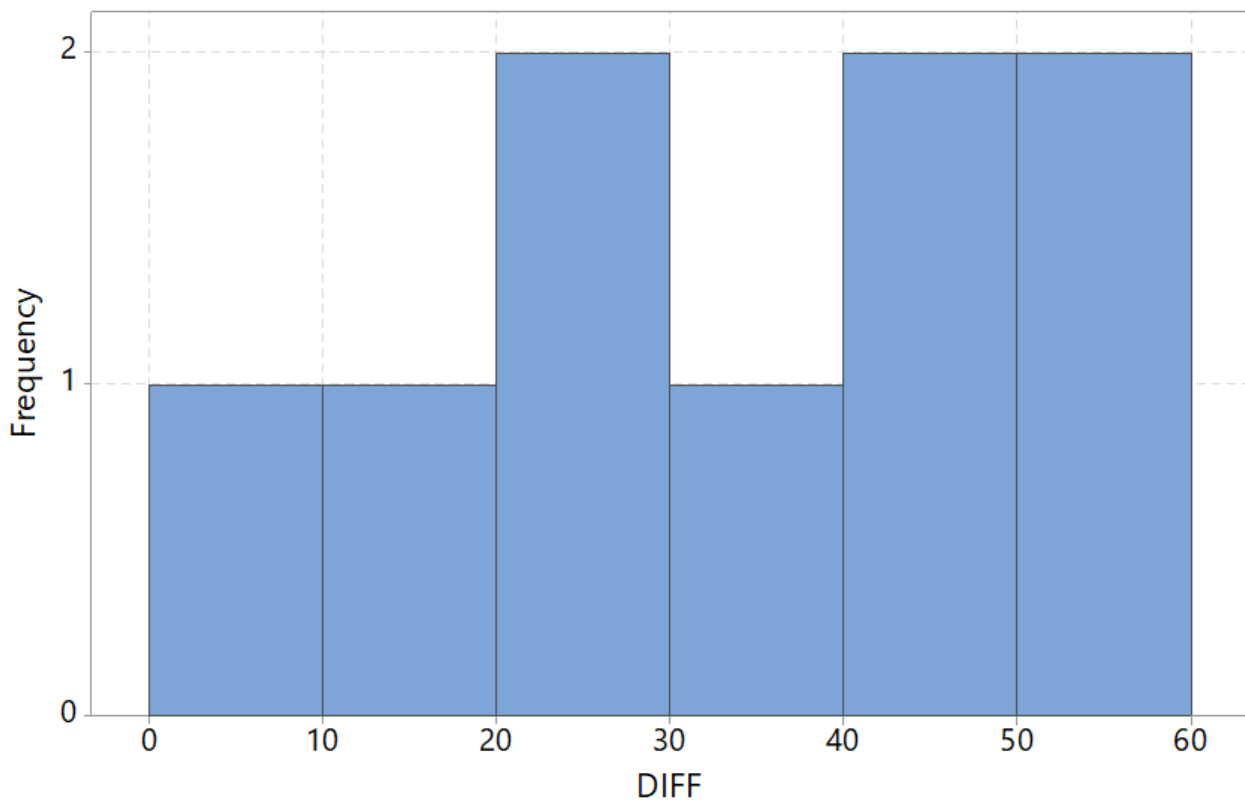
$$\text{Min} = 2$$

$$\text{Max} = 56$$

$$\sum_{i=1}^9 x_i = 282 \text{ και } \bar{x} = 31.333$$

$$s = 18.358$$

Ιστόγραμμα



Είναι ο RPULSE μεγαλύτερος του SPULSE;

- Ελέγχουμε το DIFF!

$$H_0: \mu_R = \mu_S$$

$$H_a: \mu_R > \mu_S$$

Ας κάνουμε τις υποθέσεις πιο συγκεκριμένες

$$H_0: \mu_R - \mu_S = 0$$

$$H_a: \mu_R - \mu_S > 0$$

Ας κάνουμε τις υποθέσεις ακόμα πιο συγκεκριμένες

$$H_0: \mu_{DIFF} = 0$$

$$H_a: \mu_{DIFF} > 0$$

Υπολογισμός του t statistic από το δείγμα

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{31.333 - 0}{\frac{18.358}{\sqrt{9}}} = 5.12$$

Κρίσιμες τιμές για μονόπλευρο έλεγχο t βάσει CL = 95%, $\alpha = 0.05$

$$df = 9 - 1 = 8$$

$$t_{0.05} = 1.860$$

Εφόσον

$$t = 5.12 > 1.860 = t_{0.05}$$

απορρίπτουμε τη μηδενική υπόθεση

Εναλλακτική μέθοδος με την τιμή του p (p -value) που αντιστοιχεί σε $t = 5.12$

$$p\text{-value} = 0.0005$$

Εφόσον

$$p = 0.0005 < 0.05 = \alpha$$

απορρίπτουμε τη μηδενική υπόθεση

12.2.3. Ανεξάρτητα δείγματα (independent samples)

Θα διδάξουμε τη μέθοδο του ελέγχου t για ανεξάρτητα δείγματα (independent samples) με το ακόλουθο παράδειγμα

Θεωρούμε τα ακόλουθα δύο δείγματα μισθών καθηγητών σε δημόσια (PUBLIC) και ιδιωτικά (PRIVATE) πανεπιστήμια των ΗΠΑ, που έχουν συλλεχθεί από το American Association of University Professors (σε χιλιάδες US\$)

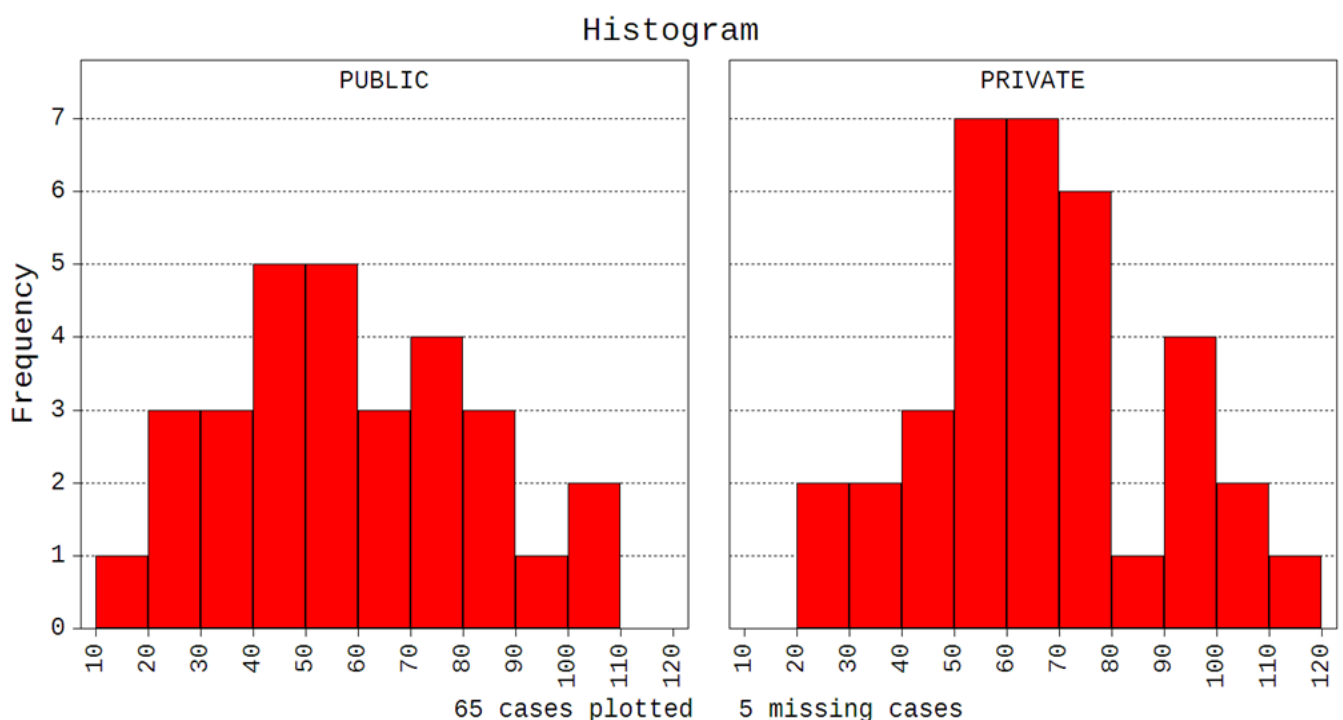
INDEX	PUBLIC	PRIVATE
1	34.2	92.9
2	90	52
3	100.4	63.1
4	24.6	118.5
5	107.4	37.7
6	63.6	102.2
7	56.8	62.9
8	41.4	53.8
9	35	101
10	54.2	68.6
11	24.4	51.5
12	56	46.4
13	58.2	45.2
14	76.8	76
15	84.2	56.1
16	79.4	77.6
17	42.2	61.6
18	81.8	78.3
19	29.2	66.3
20	15.8	31.1
21	33.8	71.1
22	40.2	73.5
23	51.2	67.6
24	41.2	52.4

25	60.2	47.2
26	88.2	59.3
27	44.6	97.5
28	64.4	27.2
29	74	81.2
30	71	24.8
31		71
32		97.3
33		92.6
34		56
35		62.3

Τα περιγραφικά στατιστικά μεγέθη των ανωτέρω μεταβλητών έχουν ως ακολούθως

Variable	N	Mean	SD	Minimum	Maximum
PUBLIC	30	57.480	23.953	15.800	107.40
PRIVATE	35	66.394	22.261	24.800	118.50

Σύγκριση ιστογραμμάτων των δύο μεταβλητών



Χρησιμοποιούμε τα ακόλουθα σύμβολα

μ_1 : μέσος μισθός καθηγητών σε δημόσια πανεπιστήμια

μ_2 : μέσος μισθός καθηγητών σε ιδιωτικά πανεπιστήμια

Τα δεδομένα συλλέχθηκαν προκειμένου να ελέγξουμε τις υποθέσεις

$$H_0: \mu_1 = \mu_2 \text{ ή } \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 \neq \mu_2 \text{ ή } \mu_1 - \mu_2 \neq 0$$

Υπενθυμίζουμε ότι η δειγματοληπτική κατανομή των μέσων όρων δειγμάτων μεγέθους n , δηλαδή η κατανομή της μεταβλητής, έχει τα εξής χαρακτηριστικά:

- είναι κανονική
- έχει μέσο όρο $\mu_{\bar{x}} = \mu$
- έχει τυπική απόκλιση ίση με $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$ (τυπικό σφάλμα)

Ε λοιπόν, η δειγματοληπτική κατανομή της διαφοράς των μέσων όρων δύο δειγμάτων μεγέθους n_1 και n_2 αντίστοιχα, δηλαδή η μεταβλητή $\bar{x}_1 - \bar{x}_2$, έχει τα ακόλουθα χαρακτηριστικά:

- είναι επίσης κανονική
- έχει μέσο όρο $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$

$$\bullet \text{ και τυπική απόκλιση } s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Για την ορθή εφαρμογή της μορφής του ελέγχου που είναι εντός ύλης και εξετάζεται εδώ, πρέπει να τηρούνται οι ακόλουθες προϋποθέσεις:

1. Τα δύο δείγματα να είναι ανεξάρτητα
2. Οι δύο πληθυσμοί από τους οποίους προέρχονται τα δείγματα να είναι κανονικά κατανεμημένοι ή τα δείγματα να είναι μεγάλα
3. Οι τυπικές αποκλίσεις των δύο πληθυσμών να είναι **περίπου** ίσες
 - Για να δούμε εάν οι τυπικές αποκλίσεις των δύο πληθυσμών είναι περίπου ίδιες, ακολουθείται ο ακόλουθος εμπειρικός κανόνας:
 - Η μεγαλύτερη από τις δύο τυπικές αποκλίσεις, να μην είναι υπερδιπλάσια (δηλαδή μεγαλύτερη από το διπλάσιο) της μικρότερης

Στατιστικό μέτρο που χρησιμοποιείται για τον έλεγχο (test statistic)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Με s_p συμβολίζουμε την κοινή τυπική απόκλιση (pooled standard deviation) που ισούται με

$$s_p = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}$$

Οι βαθμοί ελευθερίας, df , στην περίπτωση αυτού του ελέγχου ισούνται με

$$df = n_1 + n_2 - 2$$

Pooled t-Test

Purpose To perform a hypothesis test to compare two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1 The null hypothesis is $H_0: \mu_1 = \mu_2$, and the alternative hypothesis is

$$\begin{array}{lll} H_a: \mu_1 \neq \mu_2 & \text{or} & H_a: \mu_1 < \mu_2 & \text{or} & H_a: \mu_1 > \mu_2 \\ \text{(Two tailed)} & & \text{(Left tailed)} & & \text{(Right tailed)} \end{array}$$

Step 2 Decide on the significance level, α .

Step 3 Compute the value of the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{(1/n_1) + (1/n_2)}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

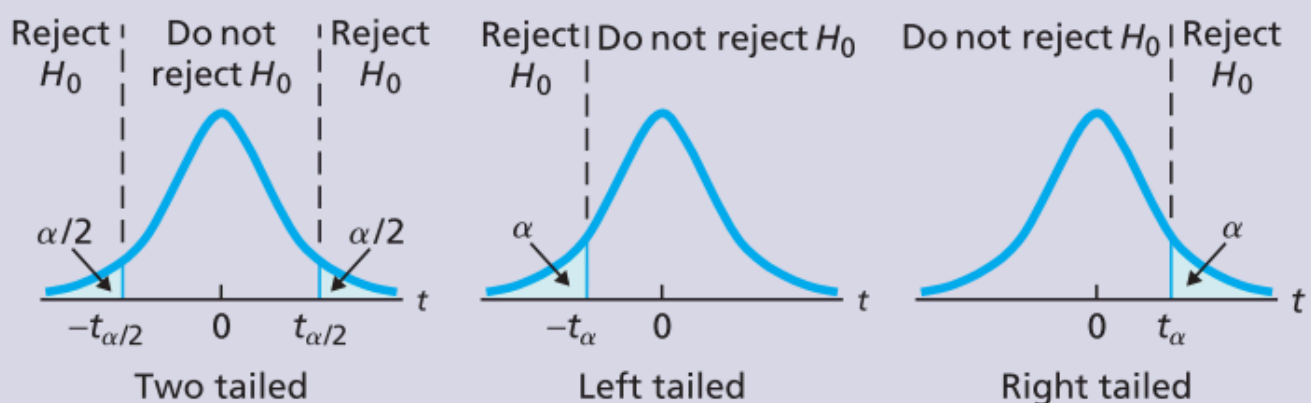
Denote the value of the test statistic t_0 .

CRITICAL-VALUE APPROACH

Step 4 The critical value(s) are

$\pm t_{\alpha/2}$ (Two tailed) or $-t_{\alpha}$ (Left tailed) or t_{α} (Right tailed)

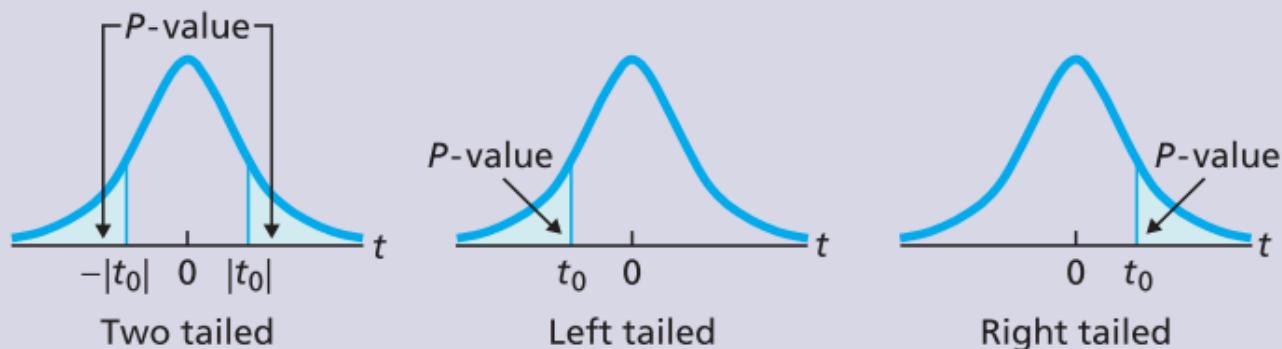
with $df = n_1 + n_2 - 2$. Use Table IV to find the critical value(s).



Step 5 If the value of the test statistic falls in the rejection region, reject H_0 ; otherwise, do not reject H_0 .

P-VALUE APPROACH

Step 4 The t -statistic has $df = n_1 + n_2 - 2$. Use Table IV to estimate the P -value, or obtain it exactly by using technology.



Step 5 If $P \leq \alpha$, reject H_0 ; otherwise, do not reject H_0 .

Βασικά στατιστικά μεγέθη του δείγματος του πρώτου πληθυσμού (public):

$$\bar{x}_1 = 57.48$$

$$s_1 = 23.95$$

$$n_1 = 30$$

Βασικά στατιστικά μεγέθη του δείγματος του δεύτερου πληθυσμού (private):

$$\bar{x}_2 = 66.39$$

$$s_2 = 22.26$$

$$n_2 = 35$$

Οι προϋποθέσεις φαίνεται να τηρούνται:

1. Τα δύο δείγματα είναι πράγματι ανεξάρτητα το ένα από το άλλο
2. Αν και δεν έχουμε γνώση των πληθυσμών, τα δύο δείγματα φαίνεται να είναι κανονικά κατανομημένα, όπως φαίνεται από τα ιστογράμματα που είδαμε προηγουμένως

- Εφόσον τα δείγματα είναι μεγάλα (δηλαδή έχουν ≥ 30 δεδομένα έκαστο), αρκεί να επιβεβαιώσουμε ότι δεν περιέχουν ακραίες τιμές (outliers)

3. Οι τυπικές αποκλίσεις των δύο δειγμάτων, 23.95 και 22.26, πράγματι μπορούμε να θεωρήσουμε ότι είναι περίπου ίσες δεδομένου ότι η μεγαλύτερη δεν είναι υπερδιπλάσια της μικρότερης

Υπολογισμός της κοινής τυπικής απόκλισης (pooled standard deviation, s_p)

$$s_p = \sqrt{\frac{(n_1-1) \cdot s_1^2 + (n_2-1) \cdot s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(30-1) \cdot 23.95^2 + (35-1) \cdot 22.26^2}{30+35-2}} = 23.053$$

Υπολογισμός του t (t-statistic)

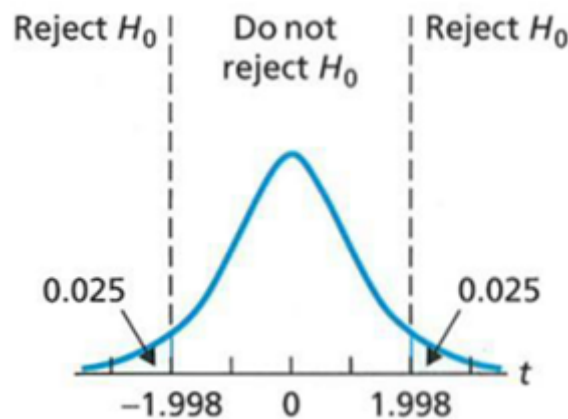
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{57.48 - 66.39}{23.053 \sqrt{\frac{1}{30} + \frac{1}{35}}} = -1.55$$

Εύρεση των κρίσιμων τιμών (για επίπεδο εμπιστοσύνης 95%)

$$df = 30 + 35 - 2 = 63$$

$$-t_{\alpha/2} = -1.998, t_{\alpha/2} = 1.998$$

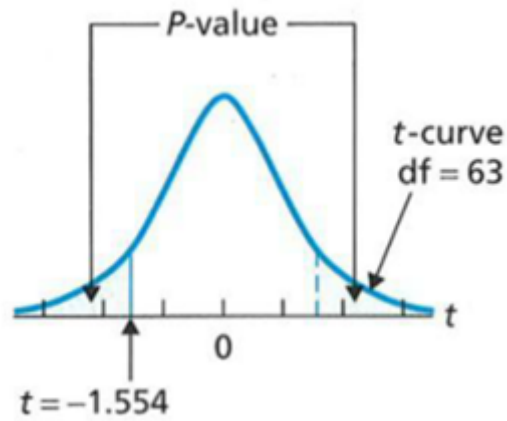
Το t που υπολογίσαμε, -1.55, πέφτει εντός της περιοχής μη απόρριψης



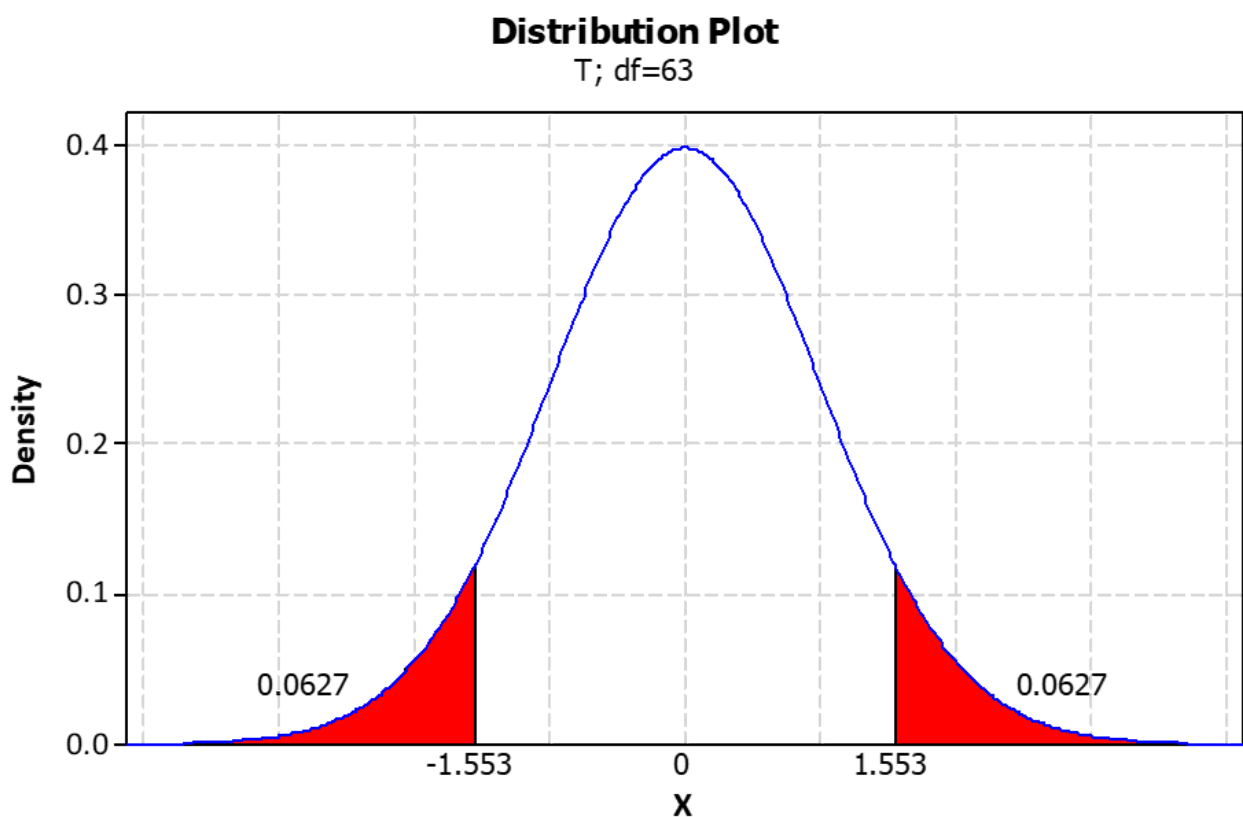
Επομένως δεν απορρίπτουμε τη μηδενική υπόθεση

- Οι μισθοί των καθηγητών δημόσιων και ιδιωτικών πανεπιστημίων των ΗΠΑ δεν διαφέρουν αρκετά για να αποφασίσουμε ότι δεν είναι ίσοι

Με τη μέθοδο της πιθανότητας (p-value), προσπαθούμε να εκτιμήσουμε από τους πίνακες της κατανομής t την p-value που αντιστοιχεί στις ουρές $-t_{\alpha/2} = -1.998$ και $t_{\alpha/2} = 1.998$



Με πρόσβαση σε στατιστικό πρόγραμμα ή διαδικτυακή εφαρμογή βρίσκουμε την ακριβή τιμή



$$p = 0.0627 + 0.0627 = 0.1252$$

Δεδομένου ότι η τιμή αυτή της πιθανότητας, $p=0.1252$, είναι μεγαλύτερη της σημαντικότητας $\alpha=0.05$ (που αντιστοιχεί σε επίπεδο εμπιστοσύνης 95%), ομοίως δεν απορρίπτουμε τη μηδενική υπόθεση

Βαθμοί πρώτης και δεύτερης προόδου εαρινού εξαμήνου 2024

ΑΜ	ΠΡΟΟΔΟΣ 1 βαθμός	ΠΡΟΟΔΟΣ 2 βαθμός
Θ12139	4.4	
Θ15008	4.4	
Θ18023	6.4	7.4

018085	6	
018171	6.4	
019126	5.1	4.8
020021	4	
020025	3.1	
020029	6	
020057	5.6	6.1
020058	7.3	
020089	7.6	
020098	4.4	
020113	4.9	
020125	6.7	2.6
020127	4	
020135	5.8	
020137	6.7	
020139	7.1	3.9
020182	3.6	4.3
021008	7.3	
021011	5.1	
021043	7.6	
021047	5.1	
021056	4.2	
021100	4	2.6
021126	7.6	6.1
021131	7.1	
021133	5.8	
021134	7.1	
021136	4.4	5.2
021138	7.3	7.4
021148	6	
021152	6.2	
021173	5.6	3
021178	6	4.3
022006	4.9	
022011	5.6	
022017	3.6	
022024	3.6	3
022027	6.7	
022030	8.2	8.7
022031	5.3	
022032	5.6	

022039	6.4	
022046	3.8	
022048	6.7	
022059	5.3	2.6
022066	5.3	
022069	5.8	2.2
022072	5.8	
022075	7.6	6.5
022078	3.8	
022090	6.9	4.3
022103	5.8	
022108	5.6	7
022119	7.3	7
022128	4.9	5.2
022131	6.4	6.1
022134	5.1	4.3
022158	1.6	
023001	6.2	
023002	8	4.8
023004	5.1	
023005	7.3	5.7
023006	6.2	3.9
023007	6	
023009	7.1	7
023010	6.9	
023011	6.9	3.5
023012	5.8	3.9
023013	6.4	1.7
023014	5.8	6.5
023015	8	6.5
023016	6.2	2.2
023017	5.1	4.8
023018	8	5.7
023019	6.9	3.9
023020	7.3	3
023022	5.1	
023023	5.1	3.9
023024	6.7	
023025	6.7	5.2
023026	4.9	
023027	6.2	4.3

023028	6.7	5.7
023029	8	4.8
023030	5.1	3
023032	7.6	3
023033	5.8	4.3
023034	5.1	3
023035	3.6	
023036	8	7
023037	6.4	4.8
023038	7.1	4.8
023039	6.9	4.3
023040	5.3	5.2
023041	5.6	6.1
023042	8	7.4
023043	7.1	5.2
023044	6.7	
023046	6.4	2.6
023047	6	2.6
023048	4.4	
023049	7.3	7.8
023050	5.8	
023051	5.1	4.8
023052	7.6	
023053	7.1	6.1
023054	6.7	
023055	7.3	3.5
023056	6	2.2
023058	5.8	6.1
023059	5.6	1.3
023060	7.1	2.6
023061	5.1	4.3
023062	7.6	4.3
023063	7.1	4.8
023064	6	2.2
023065	7.1	3.9
023066	7.8	3.9
023067	6.7	
023068	6.2	3.5
023069	7.6	6.1
023070	6.9	6.1
023071	6.2	4.3

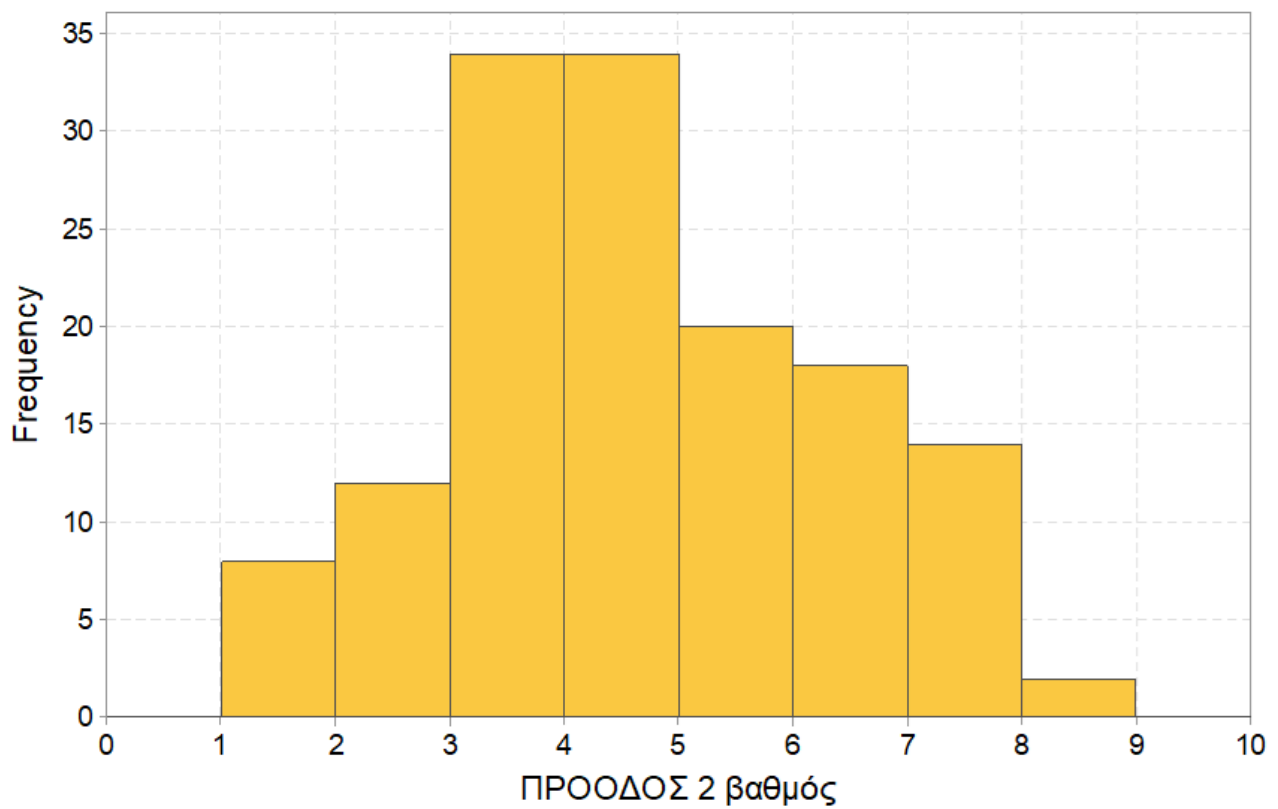
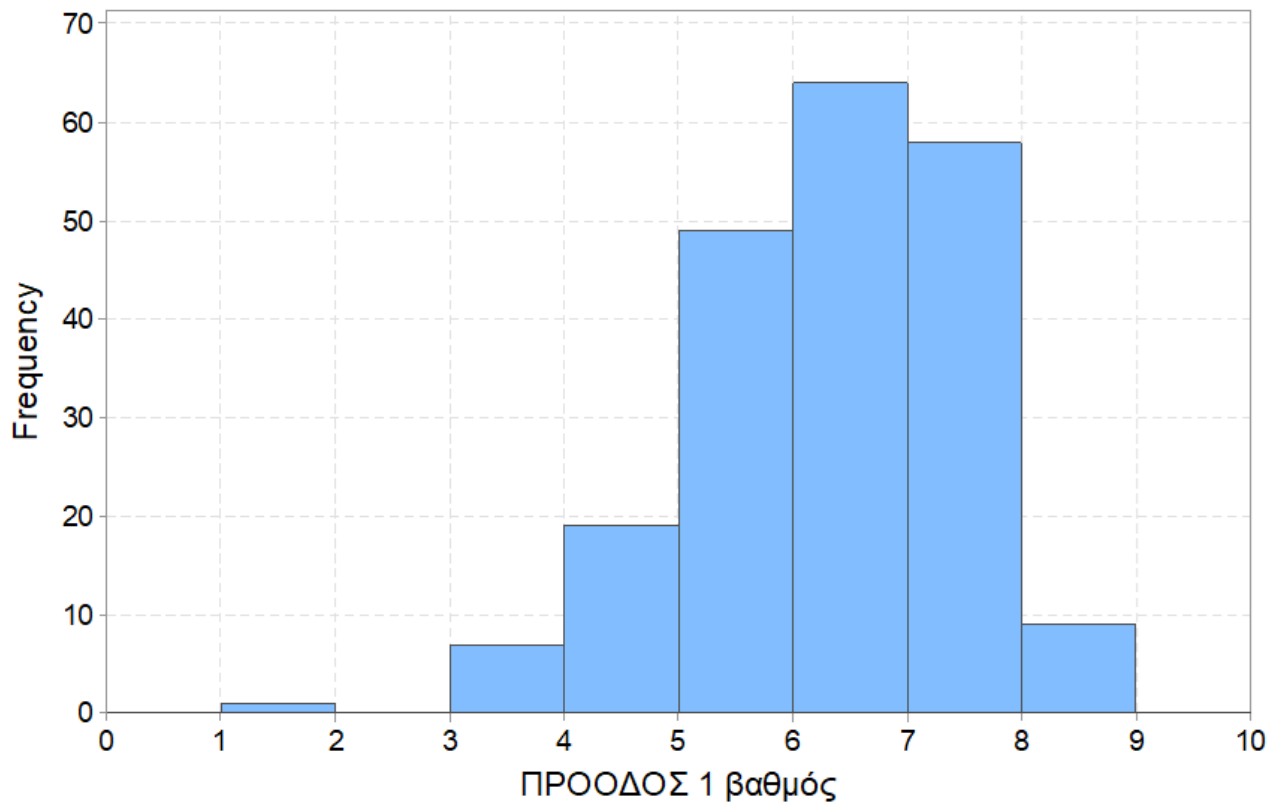
023072	6.2	3.5
023073	5.8	1.3
023074	7.1	4.3
023075	7.1	6.5
023077	6.4	5.7
023078	6.2	5.2
023079	6.9	7.8
023080	5.3	2.6
023082	7.8	
023083	6.2	5.7
023084	7.1	3.9
023085	8.2	
023086	7.1	3.9
023087	6.2	1.3
023088	7.3	6.5
023089	6.4	4.8
023090	5.3	3.5
023092	5.8	
023093	5.1	5.2
023094	7.6	4.3
023095	7.6	3.9
023096	7.8	6.1
023097	7.6	7
023098	6.4	5.7
023099	6.9	6.1
023100	6	5.2
023101	6.9	2.6
023102	7.8	7
023103	5.1	3.5
023104	6.2	4.3
023105	6.2	1.7
023107	6.9	
023108	6.9	5.7
023109		4.3
023110	7.8	3.5
023111	6.9	5.7
023112	8.4	6.5
023113	7.6	5.7
023114	4.7	5.7
023116	6	1.7
023117	6.4	4.8

023118	4.4	
023119	5.6	3.9
023120	6.2	3.5
023122	7.1	7.8
023123	5.3	
023125	7.8	3
023126	7.1	4.3
023127	5.6	3.9
023128	7.3	4.8
023130	5.1	7.8
023132	7.1	3
023133	5.3	
023134	7.3	5.7
023135	7.8	6.1
023136	4	4.3
023137	7.6	4.8
023138	7.1	
023139	5.3	
023140	7.8	3
023142	7.8	4.8
023143	7.8	5.7
023144	7.6	8.3
023146	7.8	3.9
023147	5.1	
023148	7.1	7
023149	4.9	3.9
023150	6	4.8
023151	6	1.7
023152	7.1	6.5
023153	5.3	4.3
023154	6.9	
023155	4.4	4.3
023156	5.1	3.9
023157	6.4	4.3
023158	6.9	3.9
023163	7.8	
023164		1.7
023167	4.4	4.8
023168	6.7	
023169	6	
023171	5.1	

0233076

7.8

3.5



Έλεγχος των βαθμών ως ανεξάρτητα δείγματα

- Αν και εμπεριέχονται εξαρτημένα δείγματα

- Οι φοιτητές που έδωσαν και την πρώτη και την δεύτερη πρόοδο!

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Variable	Total Count	N*	Mean	StDev	Minimum	Maximum
ΠΡΟΟΔΟΣ 1 βαθμός	209	2	6.2362	1.2011	1.6000	8.4000
ΠΡΟΟΔΟΣ 2 βαθμός	209	67	4.638	1.634	1.300	8.700

	N	Mean	StDev	SE Mean
ΠΡΟΟΔΟΣ 1 βαθμός	207	6.24	1.20	0.083
ΠΡΟΟΔΟΣ 2 βαθμός	142	4.64	1.63	0.14

Difference = μ (ΠΡΟΟΔΟΣ 1 βαθμός) - μ (ΠΡΟΟΔΟΣ 2 βαθμός)

Estimate for difference: 1.598

95% CI for difference: (1.300, 1.897)

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

T-Test of difference = 0 (vs ≠): T-Value = 10.53 P-Value = 0.000 DF = 347
Both use Pooled StDev = 1.3932

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

T-Test of difference = 0 (vs >): T-Value = 10.53 P-Value = 0.000 DF = 347
Both use Pooled StDev = 1.3932

Αν δεν είχαμε υποθέσει ίσες τυπικές αποκλίσεις

	N	Mean	StDev	SE Mean
ΠΡΟΟΔΟΣ 1 βαθμός	207	6.24	1.20	0.083
ΠΡΟΟΔΟΣ 2 βαθμός	142	4.64	1.63	0.14

Difference = μ (ΠΡΟΟΔΟΣ 1 βαθμός) - μ (ΠΡΟΟΔΟΣ 2 βαθμός)

Estimate for difference: 1.598

95% lower bound for difference: 1.333

T-Test of difference = 0 (vs >): T-Value = 9.96 P-Value = 0.000 DF = 242

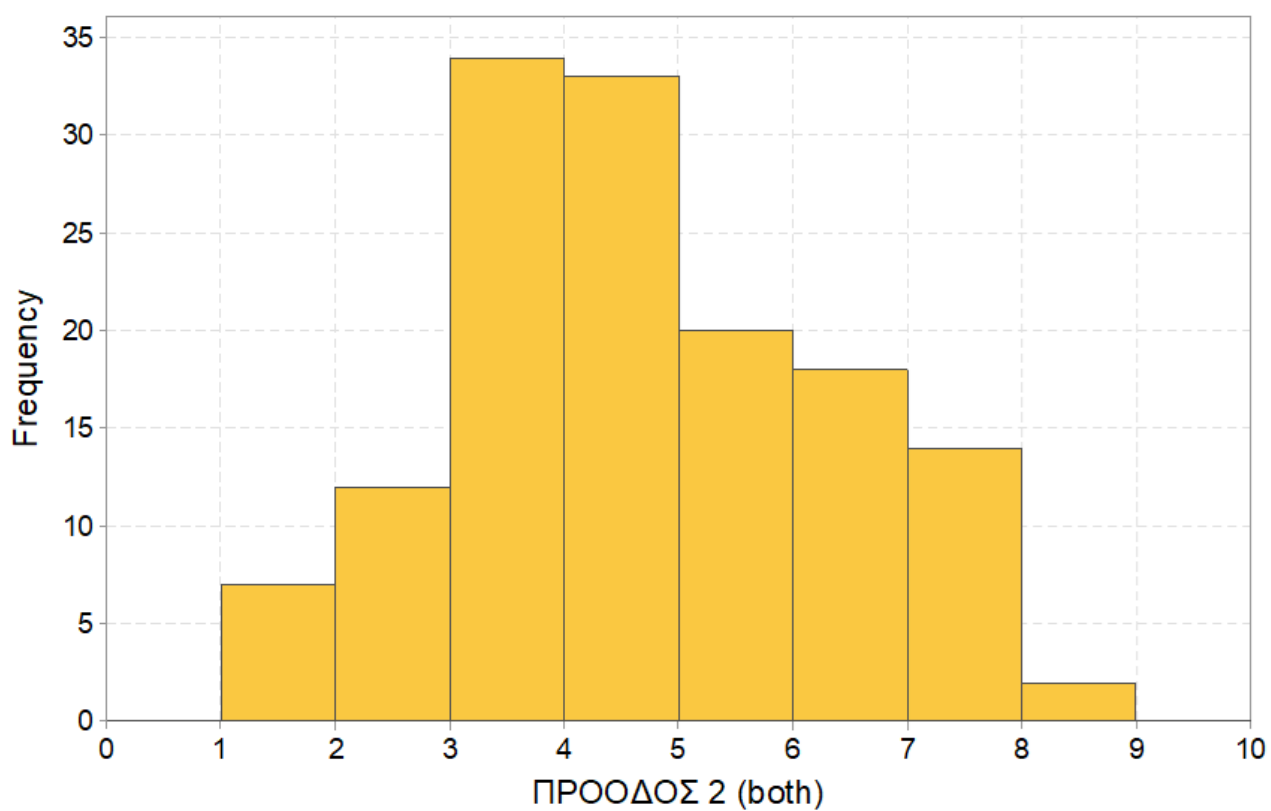
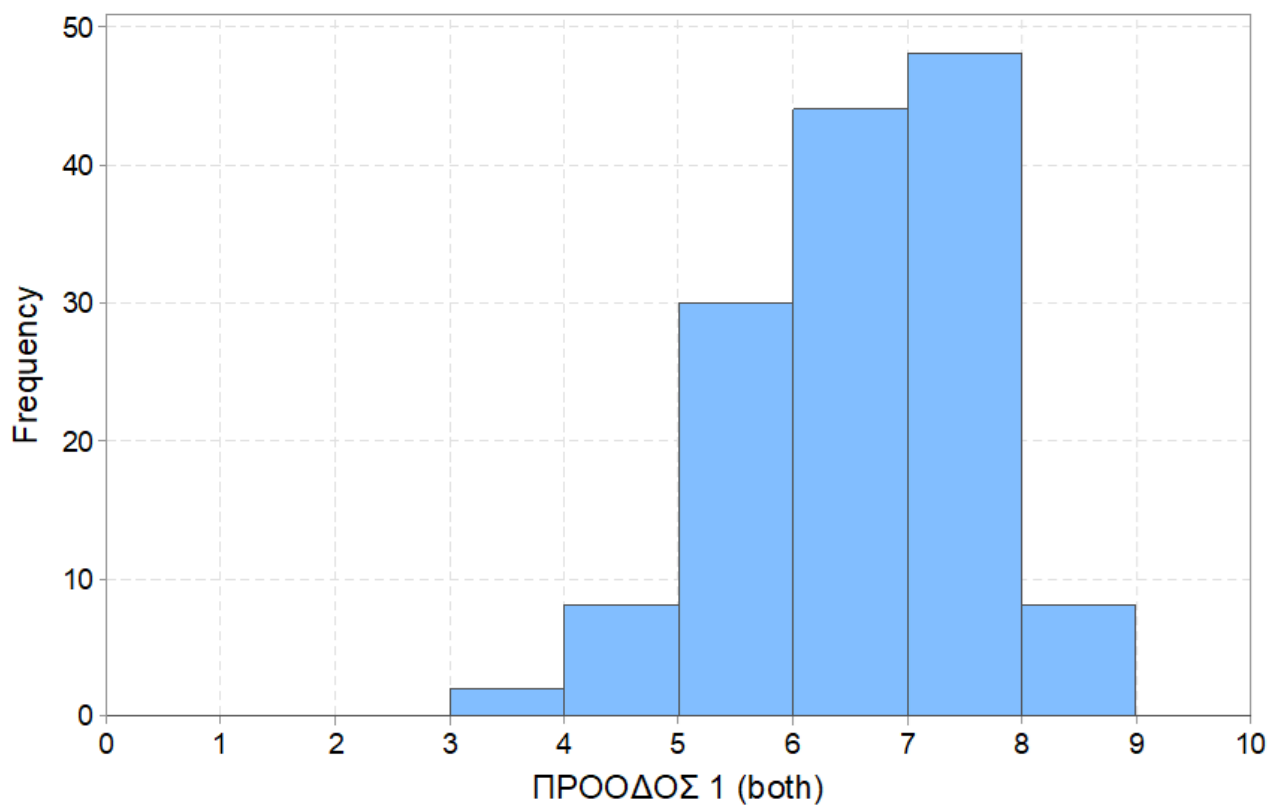
Εξετάσουμε τώρα μόνο τους φοιτητές που έδωσαν και την πρώτη και την δεύτερη πρόοδο ως εξαρτημένα δείγματα

AM (both)	ΠΡΟΟΔΟΣ 1 (both)	ΠΡΟΟΔΟΣ 2 (both)	DIFF 1-2
018023	6.4	7.4	-1
019126	5.1	4.8	0.3
020057	5.6	6.1	-0.5
020125	6.7	2.6	4.1
020139	7.1	3.9	3.2
020182	3.6	4.3	-0.7
021100	4	2.6	1.4
021126	7.6	6.1	1.5
021136	4.4	5.2	-0.8
021138	7.3	7.4	-0.1
021173	5.6	3	2.6
021178	6	4.3	1.7
022024	3.6	3	0.6
022030	8.2	8.7	-0.5
022059	5.3	2.6	2.7
022069	5.8	2.2	3.6
022075	7.6	6.5	1.1
022090	6.9	4.3	2.6
022108	5.6	7	-1.4
022119	7.3	7	0.3
022128	4.9	5.2	-0.3
022131	6.4	6.1	0.3
022134	5.1	4.3	0.8
023002	8	4.8	3.2
023005	7.3	5.7	1.6
023006	6.2	3.9	2.3
023009	7.1	7	0.1
023011	6.9	3.5	3.4
023012	5.8	3.9	1.9
023013	6.4	1.7	4.7
023014	5.8	6.5	-0.7
023015	8	6.5	1.5
023016	6.2	2.2	4
023017	5.1	4.8	0.3
023018	8	5.7	2.3
023019	6.9	3.9	3

023020	7.3	3	4.3
023023	5.1	3.9	1.2
023025	6.7	5.2	1.5
023027	6.2	4.3	1.9
023028	6.7	5.7	1
023029	8	4.8	3.2
023030	5.1	3	2.1
023032	7.6	3	4.6
023033	5.8	4.3	1.5
023034	5.1	3	2.1
023036	8	7	1
023037	6.4	4.8	1.6
023038	7.1	4.8	2.3
023039	6.9	4.3	2.6
023040	5.3	5.2	0.1
023041	5.6	6.1	-0.5
023042	8	7.4	0.6
023043	7.1	5.2	1.9
023046	6.4	2.6	3.8
023047	6	2.6	3.4
023049	7.3	7.8	-0.5
023051	5.1	4.8	0.3
023053	7.1	6.1	1
023055	7.3	3.5	3.8
023056	6	2.2	3.8
023058	5.8	6.1	-0.3
023059	5.6	1.3	4.3
023060	7.1	2.6	4.5
023061	5.1	4.3	0.8
023062	7.6	4.3	3.3
023063	7.1	4.8	2.3
023064	6	2.2	3.8
023065	7.1	3.9	3.2
023066	7.8	3.9	3.9
023068	6.2	3.5	2.7
023069	7.6	6.1	1.5
023070	6.9	6.1	0.8
023071	6.2	4.3	1.9

023072	6.2	3.5	2.7
023073	5.8	1.3	4.5
023074	7.1	4.3	2.8
023075	7.1	6.5	0.6
023077	6.4	5.7	0.7
023078	6.2	5.2	1
023079	6.9	7.8	-0.9
023080	5.3	2.6	2.7
023083	6.2	5.7	0.5
023084	7.1	3.9	3.2
023086	7.1	3.9	3.2
023087	6.2	1.3	4.9
023088	7.3	6.5	0.8
023089	6.4	4.8	1.6
023090	5.3	3.5	1.8
023093	5.1	5.2	-0.1
023094	7.6	4.3	3.3
023095	7.6	3.9	3.7
023096	7.8	6.1	1.7
023097	7.6	7	0.6
023098	6.4	5.7	0.7
023099	6.9	6.1	0.8
023100	6	5.2	0.8
023101	6.9	2.6	4.3
023102	7.8	7	0.8
023103	5.1	3.5	1.6
023104	6.2	4.3	1.9
023105	6.2	1.7	4.5
023108	6.9	5.7	1.2
023110	7.8	3.5	4.3
023111	6.9	5.7	1.2
023112	8.4	6.5	1.9
023113	7.6	5.7	1.9
023114	4.7	5.7	-1
023116	6	1.7	4.3
023117	6.4	4.8	1.6
023119	5.6	3.9	1.7
023120	6.2	3.5	2.7

023122	7.1	7.8	-0.7
023125	7.8	3	4.8
023126	7.1	4.3	2.8
023127	5.6	3.9	1.7
023128	7.3	4.8	2.5
023130	5.1	7.8	-2.7
023132	7.1	3	4.1
023134	7.3	5.7	1.6
023135	7.8	6.1	1.7
023136	4	4.3	-0.3
023137	7.6	4.8	2.8
023140	7.8	3	4.8
023142	7.8	4.8	3
023143	7.8	5.7	2.1
023144	7.6	8.3	-0.7
023146	7.8	3.9	3.9
023148	7.1	7	0.1
023149	4.9	3.9	1
023150	6	4.8	1.2
023151	6	1.7	4.3
023152	7.1	6.5	0.6
023153	5.3	4.3	1
023155	4.4	4.3	0.1
023156	5.1	3.9	1.2
023157	6.4	4.3	2.1
023158	6.9	3.9	3
023167	4.4	4.8	-0.4
0233076	7.8	3.5	4.3



Variable	Total					
	Count	N*	Mean	StDev	Minimum	Maximum
ΠΡΟΟΔΟΣ 1 (both)	140	0	6.4814	1.0696	3.6000	8.4000
ΠΡΟΟΔΟΣ 2 (both)	140	0	4.661	1.626	1.300	8.700

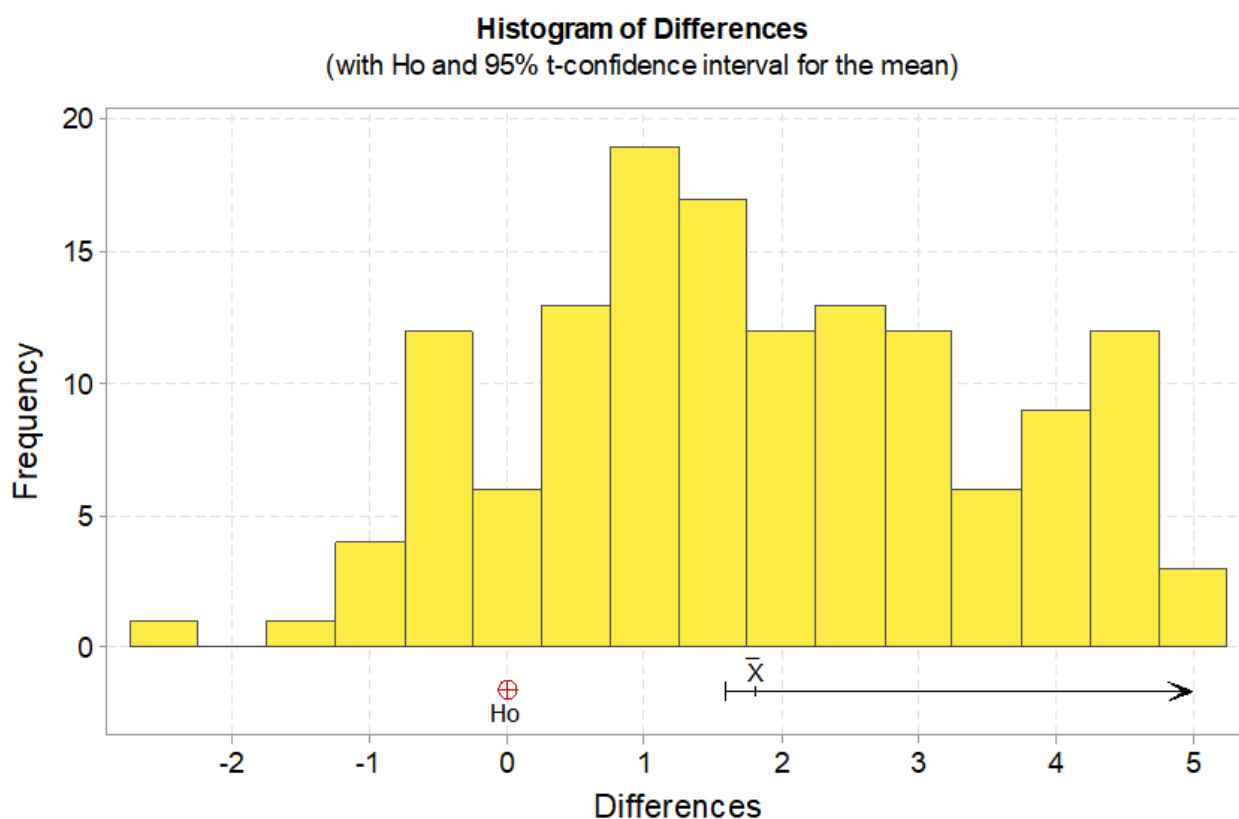
	N	Mean	StDev	SE Mean
ΠΡΟΟΔΟΣ 1 (both)	140	6.481	1.070	0.090
ΠΡΟΟΔΟΣ 2 (both)	140	4.661	1.626	0.137
Difference	140	1.820	1.638	0.138

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 > 0$$

95% lower bound for mean difference: 1.591

T-Test of mean difference = 0 (vs > 0): T-Value = 13.15 P-Value = 0.000



12.3. Διαστήματα εμπιστοσύνης και έλεγχοι ποσοστών

Εξετάσαμε

1. Δειγματοληπτικές κατανομές
2. Διαστήματα εμπιστοσύνης
3. Ελέγχους

για τον μέσο όρο ενός πληθυσμού

Τώρα εστιάζουμε σε ποσοστά (proportions), π.χ.

- % των φοιτητών του Πανεπιστημίου Πειραιώς που συνεχίζουν τις σπουδές τους σε μεταπτυχιακό επίπεδο

- % των Αθηναίων που δηλώνουν πρόθεση ψήφου υπέρ συγκεκριμένου κόμματος
- % των διδασκόντων σε ΑΕΙ που είναι γυναίκες
- % των Ελλήνων που είναι συνταξιούχοι

Επειδή τις περισσότερες φορές δεν είναι πρακτικά εφικτό να μετρήσουμε το ποσοστό ενός ολόκληρου πληθυσμού, συχνά

- Συλλέγουμε ένα δείγμα
- Μετράμε το ποσοστό στο δείγμα
- Υπολογίζουμε διαστήματα εμπιστοσύνης και διεργούμε στατιστικούς ελέγχους για το άγνωστο ποσοστό στον πληθυσμό με βάση τα δεδομένα του δείγματος

Population Proportion and Sample Proportion

Consider a population in which each member either has or does not have a specified attribute. Then we use the following notation and terminology.

Population proportion, p : The proportion (percentage) of the entire population that has the specified attribute.

Sample proportion, $\hat{p}$: The proportion (percentage) of a sample from the population that has the specified attribute.

Sample Proportion

A sample proportion, $\hat{p}$, is computed by using the formula

$$\hat{p} = \frac{x}{n},$$

where x denotes the number of members in the sample that have the specified attribute and, as usual, n denotes the sample size.

Είναι σε χρήση και το εξής σύμβολο

$$\hat{q} = 1 - \hat{p}$$

Όμως το $\hat{q}$ δεν χρησιμοποιείται στους υπολογισμούς διαστημάτων εμπιστοσύνης και ελέγχων ποσοστών

Συνήθης ορολογία

x : number of successes

$n-x$: number of failures

Correspondence between notations for means and proportions

	Parameter	Statistic
Means	μ	$\bar{x}$
Proportions	p	$\hat{p}$

Για να υπολογίσουμε διαστήματα εμπιστοσύνης και να διενεργήσουμε στατιστικούς ελέγχους, πρέπει να ξέρουμε τη δειγματοληπτική κατανομή ενός ποσοστού

The Sampling Distribution of the Sample Proportion

For samples of size n ,

- the mean of $\hat{p}$ equals the population proportion: $\mu_{\hat{p}} = p$ (i.e., the sample proportion is an unbiased estimator of the population proportion);
- the standard deviation of $\hat{p}$ equals the square root of the product of the population proportion and one minus the population proportion divided by the sample size: $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$; and
- $\hat{p}$ is approximately normally distributed for large n .

Άρα, για τη δειγματοληπτική κατανομή ενός ποσοστού $\hat{p}$ για δείγματα συγκεκριμένου μεγέθους ισχύουν τα ακόλουθα

- Ο μέσος όρος των ποσοστών όλων των δειγμάτων μεγέθους n , δηλαδή ο μέσος όρος της δειγματοληπτικής κατανομής του ποσοστού, είναι ίσος με το ποσοστό στον πληθυσμό

$$\mu_{\hat{p}} = p$$

- Η τυπική απόκλιση της δειγματοληπτικής κατανομής δειγμάτων μεγέθους n είναι

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Οι παραπάνω τύποι είναι ακριβείς όταν ισχύει ο παρακάτω εμπειρικός κανόνας

$$np \geq 5$$

$$n(1-p) \geq 5$$

Λέμε ότι το δείγμα είναι μεγάλο όταν ισχύουν και οι δύο παραπάνω συνθήκες

- Δηλαδή με τα ποσοστά, η έννοια του μεγάλου δείγματος είναι σχετική!



What Does It Mean?

- If n is large, the possible sample proportions for samples of size n have approximately a normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.

Playing Hooky From Work In Example 11.1, suppose that 19.1% of all U.S. employees play hooky, that is, that the population proportion is $p = 0.191$. Then, according to Key Fact 11.1, for samples of size 1010, the variable $\hat{p}$ is approximately normally distributed with mean $\mu_{\hat{p}} = p = 0.191$ and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.191(1-0.191)}{1010}} = 0.012.$$

12.3.1. Διάστημα εμπιστοσύνης ποσοστού

One-Proportion z-Interval Procedure

Purpose To find a confidence interval for a population proportion, p

Assumptions

1. Simple random sample
2. The number of successes, x , and the number of failures, $n - x$, are both 5 or greater.

Step 1 For a confidence level of $1 - \alpha$, use Table II to find $z_{\alpha/2}$.

Step 2 The confidence interval for p is from

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\hat{p}(1 - \hat{p})/n} \quad \text{to} \quad \hat{p} + z_{\alpha/2} \cdot \sqrt{\hat{p}(1 - \hat{p})/n},$$

where $z_{\alpha/2}$ is found in Step 1, n is the sample size, and $\hat{p} = x/n$ is the sample proportion.

Step 3 Interpret the confidence interval.

Δείγμα 1010 εργαζομένων

- 202 από αυτούς δήλωσαν ότι προφασίζονται ασθένεια για να μην πάνε στη δουλειά τους

$$n = 1010$$

$$x = 202$$

$$\hat{p} = \frac{202}{1010} = 0.2$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.2 = 0.8$$

$$n - x = 1010 - 202 = 808$$

Έλεγχος προϋποθέσεων

$$x = 202 \geq 5$$

$$n - x = 808 \geq 5$$

Συνεχίζουμε την ανάλυση

$$CL = 95\%, \alpha = 0.05$$

Κρίσιμες τιμές από τον πίνακα της τυπικής κανονικής κατανομής

$$z_{\alpha/2} = z_{0.025} = \pm 1.96$$

Υπολογισμός του διαστήματος εμπιστοσύνης

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)$$

$$= \left(0.2 - 1.96 \cdot \sqrt{\frac{0.2 \cdot (1-0.2)}{1010}}, 0.2 + 1.96 \cdot \sqrt{\frac{0.2 \cdot (1-0.2)}{1010}} \right)$$

$$= (0.175, 0.225)$$

Έλεγχος με το Minitab

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	202	1010	0.200000	(0.175331, 0.224669)

Using the normal approximation.

Margin of Error for the Estimate of p

The **margin of error** for the estimate of p is

$$E = z_{\alpha/2} \cdot \sqrt{\hat{p}(1 - \hat{p})/n}$$

Το περιθώριο σφάλματος λοιπόν (που συχνά αναφέρεται στις δημοσκοπήσεις) είναι

$$1.96 \cdot \sqrt{\frac{0.2 \cdot (1-0.2)}{1010}} = 0.025$$

Πόσο μεγάλο πρέπει να είναι ένα δείγμα για να εξασφαλίσουμε ένα συγκεκριμένο περιθώριο σφάλματος;

- Αυτό είναι ένα συνηθισμένο ζητούμενο στις δημοσκοπήσεις

Sample Size for Estimating p

A $(1 - \alpha)$ -level confidence interval for a population proportion that has a margin of error of at most E can be obtained by choosing

$$n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2$$

rounded up to the nearest whole number. If you can make an educated guess, $\hat{p}_g$ (g for guess), for the observed value of $\hat{p}$, then you should instead choose

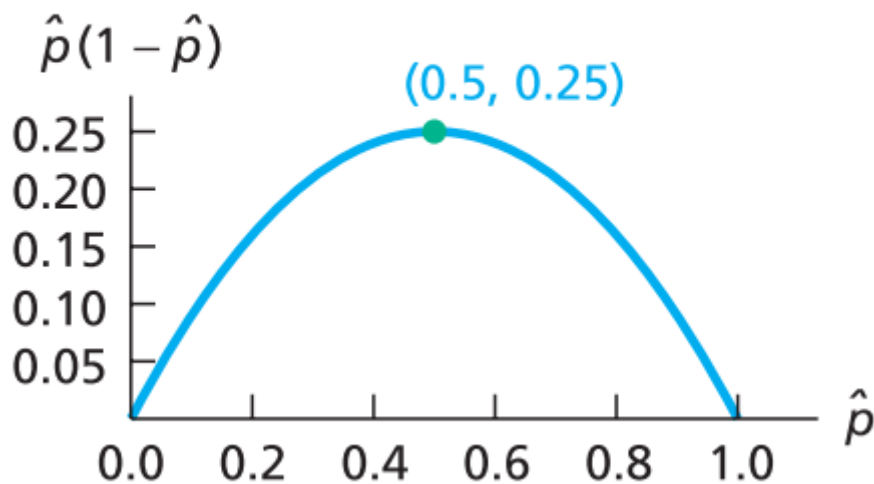
$$n = \hat{p}_g(1 - \hat{p}_g) \left(\frac{z_{\alpha/2}}{E} \right)^2$$

rounded up to the nearest whole number.

Βέβαια, συνήθως δεν γνωρίζουμε την τιμή του ποσοστού!

- Αν όμως μπορούμε να προβλέψουμε την κατά προσέγγιση τιμή του ποσοστού, τότε κάνουμε χρήση του δεύτερου τύπου, που περιέχει και την κατ' εκτίμηση τιμή του ποσοστού
- Ο δεύτερος τύπος δίνει πιο ακριβή αποτελέσματα

Graph of $\hat{p}(1 - \hat{p})$ versus $\hat{p}$



Το παραπάνω σχήμα δείχνει ότι χρειαζόμαστε μεγαλύτερο δείγμα για να προβλέψουμε τιμές ποσοστού που είναι πλησιέστερα στο 50% παρά σε μικρές (ή μεγάλες) τιμές

- Οπότε καταλαβαίνουμε ότι χρειαζόμαστε μεγαλύτερο δείγμα για να προβλέψουμε το ποσοστό των κομμάτων εκείνων που έχουν ποσοστό γύρω στο 50%!

Ας δούμε τι μας δίνει ο πρώτος (απλούστερος) τύπος για διαφορετικές τιμές περιθωρίου σφάλματος (σε επίπεδο εμπιστοσύνης 95%)

$$n = 0.25 \left(\frac{1.96}{0.01} \right)^2 = 9604$$

$$n = 0.25 \left(\frac{1.96}{0.05} \right)^2 = 384.16 \Rightarrow 385$$

$$n = 0.25 \left(\frac{1.96}{0.1} \right)^2 = 96.04 \Rightarrow 97$$

Δηλαδή

- Για να υπολογίσουμε διάστημα εμπιστοσύνης 95% με περιθώριο λάθος 0.01 (δηλαδή 1%), χρειαζόμαστε δείγμα μεγέθους 9604
 - Βλέπουμε γιατί τα exit poll συνήθως βασίζονται σε τόσο μεγάλα δείγματα
- Για να υπολογίσουμε διάστημα εμπιστοσύνης 95% με περιθώριο λάθος 0.05 (δηλαδή 5%), χρειαζόμαστε δείγμα μεγέθους 385
- Για να υπολογίσουμε διάστημα εμπιστοσύνης 95% με περιθώριο λάθος 0.1 (δηλαδή 10%), χρειαζόμαστε δείγμα μεγέθους 97

12.3.2. Διάστημα εμπιστοσύνης ποσοστού με το μέθοδο plus four **[ΕΚΤΟΣ ΥΛΗΣ ΓΙΑ ΕΞΕΤΑΣΗ 2025-06]**

Θα εξετάσουμε τη μέθοδο plus four για τον υπολογισμό διαστημάτων εμπιστοσύνης

- Είναι πιο ακριβής για μικρά δείγματα

Σε δείγμα 60 φοιτητών ενός πανεπιστημίου, οι 32 επέλεξαν το μάθημα των μαθηματικών

(<https://vrcacademy.com/tutorials/plus-four-confidence-interval-proportion-example/>)

Να υπολογιστεί διάστημα εμπιστοσύνης για το ποσοστό των φοιτητών όλου του πανεπιστημίου που επέλεξαν το μαθηματικά

$$n = 60$$

$$x = 32$$

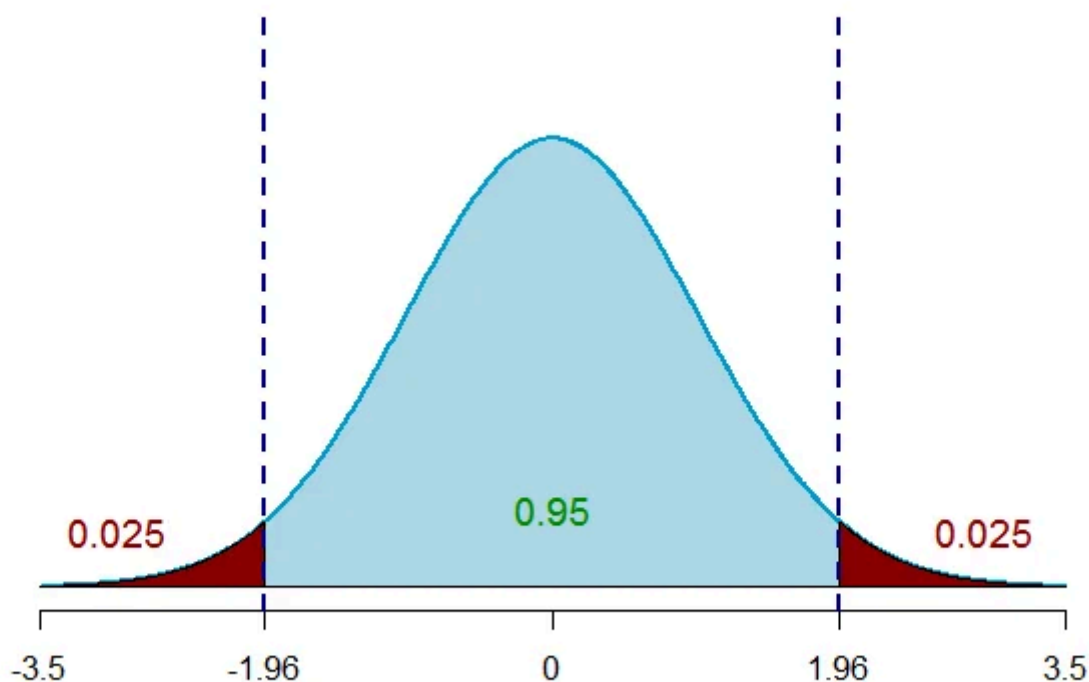
$$\hat{p} = \frac{x}{n} = \frac{32}{60} = 0.533$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.533 = 0.467$$

$$CL = 95\%, \alpha = 0.05$$

Εύρεση κρίσιμων τιμών από τον πίνακα της τυποποιημένης κανονικής κατανομής

Z-Critical values



$$z_{\alpha/2} = z_{0.025} = \pm 1.96$$

Υπολογισμός του διαστήματος εμπιστοσύνης

$$\begin{aligned} & \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \\ &= \left(0.533 - 1.96 \cdot \sqrt{\frac{0.533 \cdot (1-0.533)}{60}}, 0.533 + 1.96 \cdot \sqrt{\frac{0.533 \cdot (1-0.533)}{60}} \right) \\ &= (0.531 - 0.126, 0.531 + 0.126) \\ &= (0.405, 0.657) \end{aligned}$$

Τώρα με τη μέθοδο plus four

- Προσθέτουμε 2 στην τιμή του x
- Προσθέτουμε 4 στο μέγεθος του δείγματος (n)

$$n = 60$$

$$x = 32$$

$$\hat{p} = \frac{x+2}{n+4} = \frac{32+2}{60+4} = \frac{34}{64} = 0.531$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.531 = 0.469$$

$$CL = 95\%, \alpha = 0.05$$

Κρίσιμες τιμές από τον πίνακα της τυποποιημένης κανονικής κατανομής

$$z_{\alpha/2} = z_{0.025} = \pm 1.96$$

Υπολογισμός του διαστήματος εμπιστοσύνης

$$\begin{aligned} & \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n+4}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n+4}} \right) \\ &= \left(0.531 - 1.96 \cdot \sqrt{\frac{0.531 \cdot (1-0.531)}{60+4}}, 0.531 + 1.96 \cdot \sqrt{\frac{0.531 \cdot (1-0.531)}{60+4}} \right) \\ &= (0.531 - 0.122, 0.531 + 0.122) \\ &= (0.409, 0.653) \end{aligned}$$

Άλλο ένα παράδειγμα υπολογισμού διαστήματος εμπιστοσύνης ποσοστού με τη μέθοδο plus four

(<https://pressbooks.lib.vt.edu/introstatistics/chapter/behavior-of-confidence-intervals-for-a-proportion/>)

Σε δείγμα 25 φοιτητών ενός πανεπιστημίου, οι 6 δηλώνουν ότι καπνίζουν

Να υπολογιστεί διάστημα εμπιστοσύνης για το ποσοστό των φοιτητών που καπνίζουν σε όλο το πληθυσμό των φοιτητών του πανεπιστημίου

$$n = 25$$

$$x = 6$$

$$\hat{p} = \frac{x+2}{n+4} = \frac{6+2}{25+4} = \frac{8}{29} = 0.276$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.276 = 0.724$$

$$CL = 95\%, \alpha = 0.05$$

Κρίσιμες τιμές από τον πίνακα της τυποποιημένης κανονικής κατανομής

$$z_{\alpha/2} = z_{0.025} = \pm 1.96$$

Υπολογισμός του διαστήματος εμπιστοσύνης

$$\begin{aligned} & \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n+4}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n+4}} \right) \\ &= \left(0.276 - 1.96 \cdot \sqrt{\frac{0.276 \cdot (1-0.276)}{25+4}}, 0.276 + 1.96 \cdot \sqrt{\frac{0.276 \cdot (1-0.276)}{25+4}} \right) \\ &= (0.276 - 0.163, 0.276 + 0.163) \\ &= (0.113, 0.439) \end{aligned}$$

12.3.3. Έλεγχος ποσοστού

One-Proportion z-Test

Purpose To perform a hypothesis test for a population proportion, p

Assumptions

1. Simple random sample
2. Both np_0 and $n(1 - p_0)$ are 5 or greater

Step 1 The null hypothesis is $H_0: p = p_0$, and the alternative hypothesis is

$$\begin{array}{ccc} H_a: p \neq p_0 & \text{or} & H_a: p < p_0 & \text{or} & H_a: p > p_0 \\ \text{(Two tailed)} & & \text{(Left tailed)} & & \text{(Right tailed)} \end{array}$$

Step 2 Decide on the significance level, α .

Step 3 Compute the value of the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

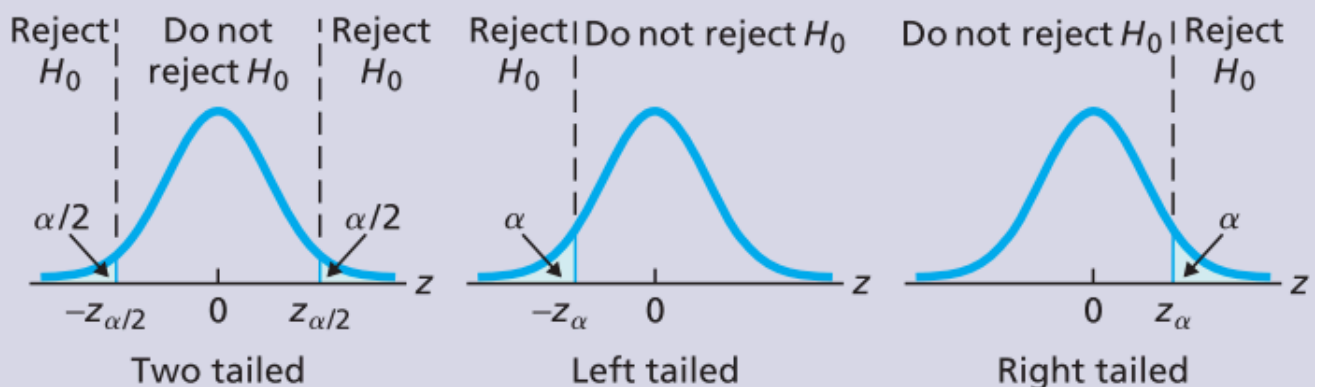
and denote that value z_0 .

CRITICAL-VALUE APPROACH

Step 4 The critical value(s) are

$\pm z_{\alpha/2}$ (Two tailed) or $-z_{\alpha}$ (Left tailed) or z_{α} (Right tailed)

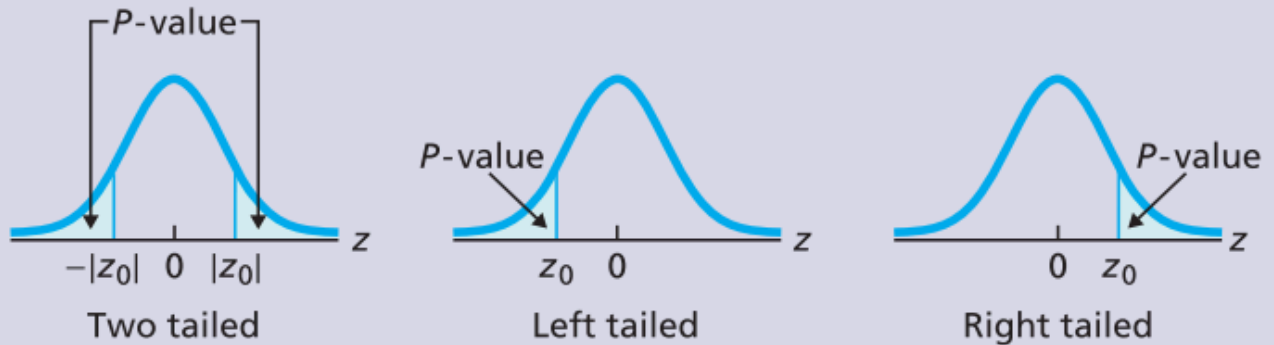
Use Table II to find the critical value(s).



Step 5 If the value of the test statistic falls in the rejection region, reject H_0 ; otherwise, do not reject H_0 .

P-VALUE APPROACH

Step 4 Use Table II to obtain the P -value.



Step 5 If $P \leq \alpha$, reject H_0 ; otherwise, do not reject H_0 .

Δημοσκόπηση 1053 ενηλίκων, 548 από αυτούς δήλωσαν ότι στηρίζουν μια συγκεκριμένη οικονομική πολιτική

Σε επίπεδο εμπιστοσύνης 95%, ισχύει ότι την πολιτική τη στηρίζει τουλάχιστον το 50% του πληθυσμού (δηλαδή 50% ή και παραπάνω);

Στήνουμε το ζεύγος υποθέσεων

$$H_0: p = 0.50 = p_0$$

$$H_a: p > 0.50 = p_0$$

$$p_0 = 0.50 = 50\%$$

$$n = 1053$$

$$x = 548$$

$$\hat{p} = \frac{548}{1053} = 0.52 = 52\%$$

Οι προϋποθέσεις πληρούνται

$$np_0 = 1053 \times 0.5 = 526.5 > 5$$

$$n(1-p_0) = 1053 \times (1-0.5) = 526.5 > 5$$

Επίπεδο εμπιστοσύνης (CL) και σημαντικότητα (α)

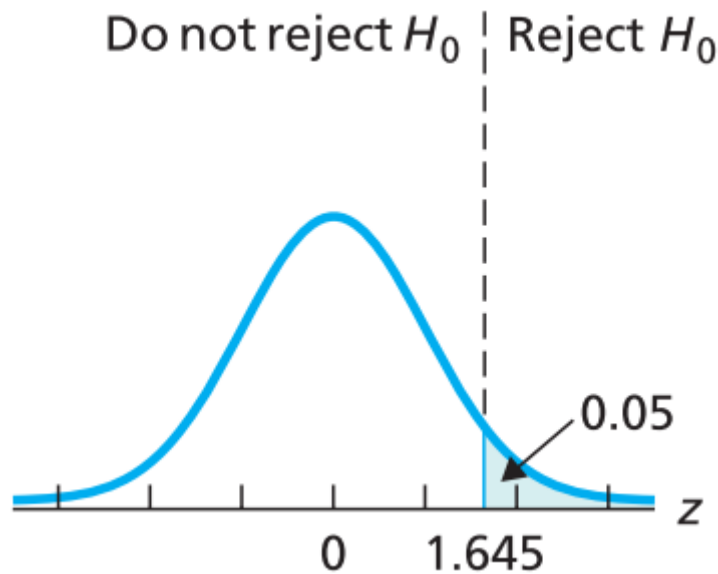
$$CL = 95\%, \alpha = 0.05$$

Υπολογίζουμε το στατιστικό μέγεθος

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

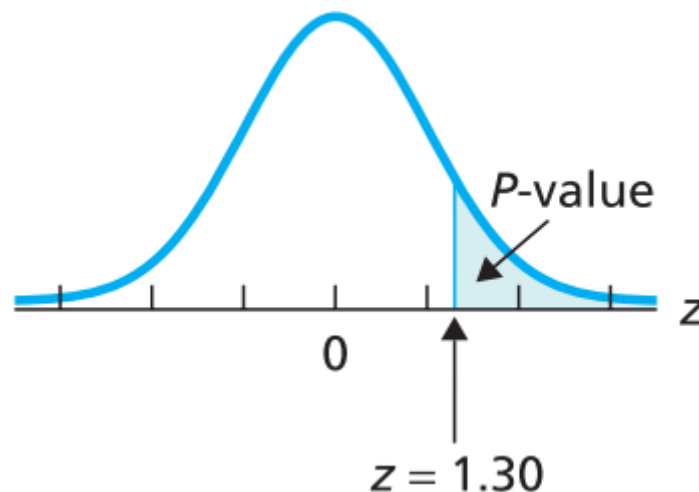
$$z = \frac{0.520 - 0.50}{\sqrt{(0.50)(1 - 0.50)/1053}} = 1.30$$

Έλεγχος με τη μέθοδο των κρίσιμων τιμών

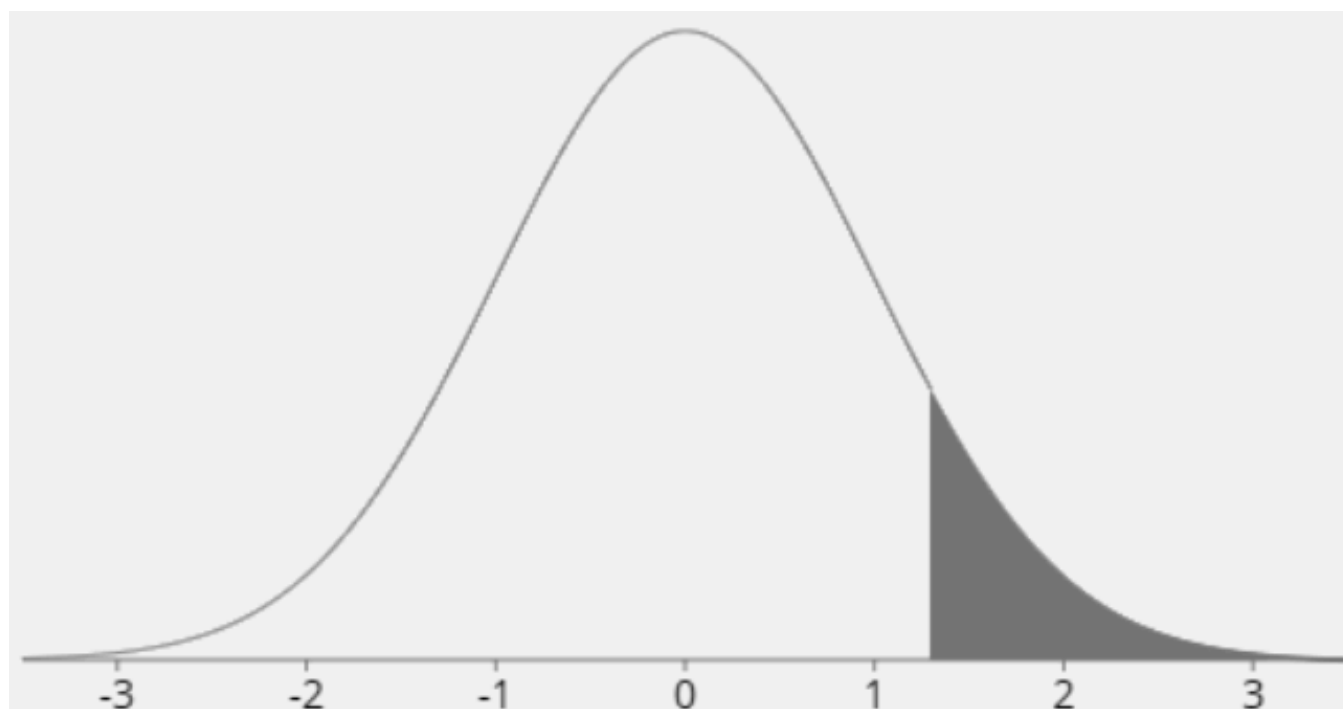


$$z = 1.30 < 1.645 = z_\alpha \Rightarrow \text{δεν απορρίπτουμε την } H_0$$

Έλεγχος με τη μέθοδο του p-value



$$p = 0.0968 > 0.05 = \alpha \Rightarrow \text{δεν απορρίπτουμε την } H_0$$



Specify Parameters:

Mean
SD

- ☒ Above
☐ Below
☐ Between and
☐ Outside and

Results:

Area (probability) =

Έλεγχος με το Minitab

Test of $p = 0.5$ vs $p > 0.5$

Sample	X	N	Sample p	95% Lower Bound	Z-Value	P-Value
1	548	1053	0.520418	0.495095	1.33	0.093

12.4. Έλεγχοι χ^2

Οι έλεγχοι που βασίζονται στην κατανομή χ^2 δεν αφορούν πληθυσμιακές παραμέτρους όπως το μέσο όρο

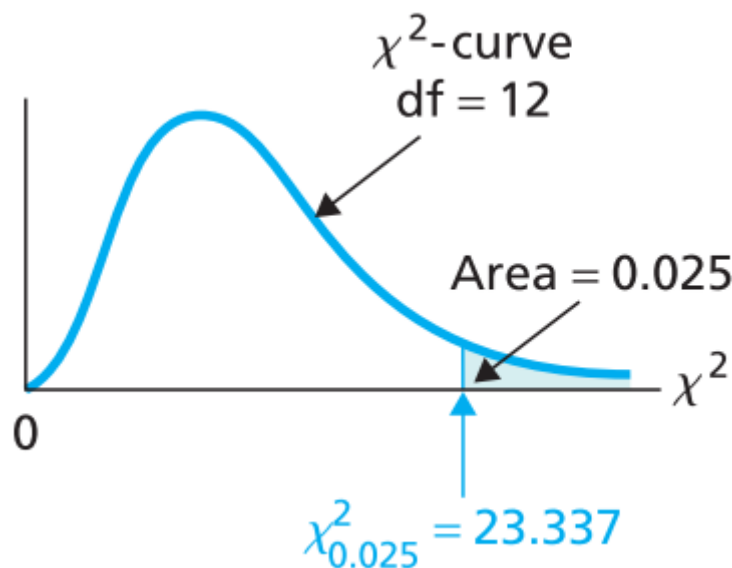
- Σε αντίθεση με τους προηγούμενους ελέγχους που εξετάσαμε

Έλεγχοι που εμπίπτουν σε αυτή την κατηγορία

1. Έλεγχος προσαρμογής χ^2 (chi-square goodness of fit ή GOF test)
2. Έλεγχοι ανεξαρτησίας χ^2 για πίνακες ενδεχομένων (contingency tables)
3. Έλεγχος ομοιογένειας χ^2 (chi-square homogeneity test)

Οι έλεγχοι που βασίζονται στην κατανομή χ^2 είναι πάντα μονόπλευροι και μάλιστα δεξιόστροφοι

Χρήση πινάκων κατανομής χ^2



12.4.1. Έλεγχος προσαρμογής

Έλεγχος προσαρμογής χ^2 (chi-square goodness of fit ή GOF)

Μια δημοσκόπηση δείγματος 1000 ψηφοφόρων, προέβλεπε τα ακόλουθα αποτελέσματα για τις βουλευτικές εκλογές, σε σχέση με τα πραγματικά αποτελέσματα των εκλογών

Κόμμα	Δημοσκόπηση (άτομα)	Δημοσκόπηση (%)	Αποτελέσματα εκλογών (%)
Δεξιά	341	34.1	36

Σοσιαλιστές	255	25.5	26
Κέντρο	153	15.3	15
Φιλελεύθεροι	77	7.7	7
Κομμουνιστές	96	9.6	9
Ανεξάρτητοι	78	7.8	7
ΣΥΝΟΛΟ	1000	100	100

Θα ελέγξουμε κατά πόσον η δημοσκόπηση προέβλεψε με αρκετή ακρίβεια τα αποτελέσματα των εκλογών με τη μέθοδο του ελέγχου προσαρμογής (**Goodness of Fit** ή **GOF**) χ^2

Με τη μέθοδο αυτή συγκρίνουμε τις πραγματικές/παρατηρούμενες (**observed**) συχνότητες της δημοσκόπησης, με τις αναμενόμενες (**expected**) συχνότητες βάσει των αποτελεσμάτων των εκλογών, τις οποίες υπολογίζουμε και παραθέτουμε στην επόμενο πίνακα

- Εκφράζουμε τα αποτελέσματα των εκλογών όχι ως ποσοστά τοις εκατό (%) αλλά ως ποσοστά της μονάδας (p)
 - Για παράδειγμα, το 36% γίνεται 0.36, ενώ το 9% γίνεται 0.09

Κόμμα	Πραγματικές συχνότητες (Observed ή O)	Αποτελέσματα εκλογών (p)	Αναμενόμενες συχνότητες (Expected ή E)
Δεξιά	341	0.36	360
Σοσιαλιστές	255	0.26	260
Κέντρο	153	0.15	150
Φιλελεύθεροι	77	0.07	70
Κομμουνιστές	96	0.09	90
Ανεξάρτητοι	78	0.07	70
ΣΥΝΟΛΟ	1000	1	1000

Για να υπολογίσουμε τις αναμενόμενες συχνότητες (E) πολλαπλασιάσαμε το συνολικό μέγεθος του δείγματος ($n = 1000$) με το ποσοστό που πήρε κάθε κόμμα στις εκλογές (p)

$$E_{\text{Δεξιά}} = n \times p_{\text{Δεξιά}} = 1000 \times 0.36 = 360$$

$$E_{\text{Σοσιαλιστές}} = n \times p_{\text{Σοσιαλιστές}} = 1000 \times 0.26 = 260$$

$$E_{\text{Κέντρο}} = n \times p_{\text{Κέντρο}} = 1000 \times 0.15 = 150$$

$$E_{\text{Φιλελεύθεροι}} = n \times p_{\text{Φιλελεύθεροι}} = 1000 \times 0.07 = 70$$

$$E_{\text{Κομμουνιστές}} = n \times p_{\text{Κομμουνιστές}} = 1000 \times 0.09 = 90$$

$$E_{\text{Ανεξάρτητοι}} = n \times p_{\text{Ανεξάρτητοι}} = 1000 \times 0.07 = 70$$

Expected Frequencies for a Goodness-of-Fit Test

In a chi-square goodness-of-fit test, the expected frequency for each possible value of the variable is found by using the formula

$$E = np,$$

where n is the sample size and p is the relative frequency (or probability) given for the value in the null hypothesis.

Εάν η δημοσκόπηση ήταν απολύτως ακριβής, τότε στο δείγμα των 1000 πολιτών που ερωτήθηκαν, “Δεξιά” θα έπρεπε να είχαν δηλώσει ότι ψηφίζουν οι 360 (και όχι οι 341), “Σοσιαλιστές” θα έπρεπε να είχαν δηλώσει ότι ψηφίζουν οι 260 (και όχι οι 255), κλπ.

Το ερώτημα είναι κατά πόσον οι διαφορές σε αυτά που προέβλεψε η δημοσκόπηση και τα πραγματικά αποτελέσματα των εκλογών είναι στατιστικά σημαντικές

Οι προϋποθέσεις που πρέπει να τηρούνται και τα πλήρη βήματα του ελέγχου έχουν ως εξής

Chi-Square Goodness-of-Fit Test

Purpose To perform a hypothesis test for the distribution of a variable

Assumptions

1. All expected frequencies are 1 or greater
2. At most 20% of the expected frequencies are less than 5
3. Simple random sample

Step 1 The null and alternative hypotheses are, respectively,

H_0 : The variable has the specified distribution

H_a : The variable does not have the specified distribution.

Step 2 Decide on the significance level, α .

Step 3 Compute the value of the test statistic

$$\chi^2 = \sum (O - E)^2 / E,$$

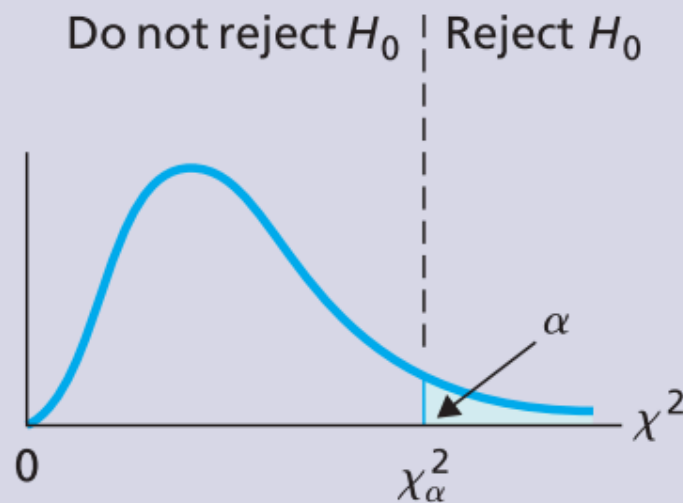
where O and E represent observed and expected frequencies, respectively. Denote the value of the test statistic χ_0^2 .

Ο έλεγχος ολοκληρώνεται είτε με τη μέθοδο των κρίσιμων τιμών είτε με τη μέθοδο της p-value

- **ΠΡΟΣΟΧΗ:** Ο έλεγχος προσαρμογής χ^2 είναι πάντα μονόπλευρος (και μάλιστα δεξιόστροφος)

CRITICAL-VALUE APPROACH

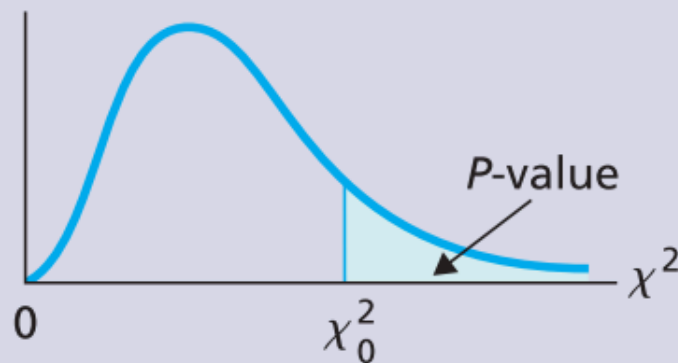
Step 4 The critical value is χ^2_α with $df = c - 1$, where c is the number of possible values for the variable. Use Table V to find the critical value.



Step 5 If the value of the test statistic falls in the rejection region, reject H_0 ; otherwise, do not reject H_0 .

P-VALUE APPROACH

Step 4 The χ^2 -statistic has $df = c - 1$, where c is the number of possible values for the variable. Use Table V to estimate the P -value, or obtain it exactly by using technology.



Step 5 If $P \leq \alpha$, reject H_0 ; otherwise, do not reject H_0 .

Οι προϋποθέσεις τηρούνται

- Όλες οι αναμενόμενες συχνότητες είναι πολύ μεγαλύτερες του 1
- Καμία από τις αναμενόμενες συχνότητες δεν είναι μικρότερη του 5
- Θεωρούμε ότι το δείγμα της δημοσκόπησης επιλέχθηκε με τυχαίο τρόπο

Μηδενική και εναλλακτική υπόθεση

H_0 : Τα ποσοστά της δημοσκόπησης ακολουθούν την κατανομή των αποτελεσμάτων των εκλογών

H_a : Τα ποσοστά της δημοσκόπησης **δεν** ακολουθούν την κατανομή των αποτελεσμάτων των εκλογών

Επίπεδο εμπιστοσύνης

$$CL = 95\%, \alpha = 0.05$$

Υπολογισμός της τιμής του χ^2

Distribution of the χ^2 -Statistic for a Goodness-of-Fit Test

For a chi-square goodness-of-fit test, the test statistic

$$\chi^2 = \sum (O - E)^2 / E$$

has approximately a chi-square distribution if the null hypothesis is true. The number of degrees of freedom is 1 less than the number of possible values for the variable under consideration.

Κόμμα	Observed (O)	Expected (E)	(O-E)	(O-E) ²	χ^2 (O-E) ² /E
Δεξιά	341	360	-19	361	1.003
Σοσιαλιστές	255	260	-5	25	0.096
Κέντρο	153	150	3	9	0.060
Φιλελεύθεροι	77	70	7	49	0.700
Κομμουνιστές	96	90	6	36	0.400
Ανεξάρτητοι	78	70	8	64	0.914
ΣΥΝΟΛΟ	1000	1000	0		3.173

$$\chi^2 = 3.173$$

Με τη μέθοδο των κρίσιμων τιμών

Βαθμοί ελευθερίας: $df = c - 1$, όπου c ο αριθμός των διαθέσιμων επιλογών (δηλαδή των κομμάτων)

$$\text{Οπότε } df = 6 - 1 = 5$$

Εύρεση κρίσιμης τιμής από τον πίνακα της κατανομής χ^2 για 5 βαθμούς ελευθερίας, α ίσο με 0.05 και μονόπλευρο έλεγχο

$\chi^2_{0.10}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	df
2.706	3.841	5.024	6.635	7.879	1
4.605	5.991	7.378	9.210	10.597	2
6.251	7.815	9.348	11.345	12.838	3
7.779	9.488	11.143	13.277	14.860	4
9.236	11.070	12.833	15.086	16.750	5

Επειδή $\chi^2 = 3.173 < 11.070 = \chi^2_{critical}$, δεν απορρίπτουμε την μηδενική υπόθεση

Για τη μέθοδο της p-value

- Από τον παραπάνω πίνακα, εφόσον το χ^2 που βρήκαμε είναι 3.173, και επομένως είναι μικρότερο ακόμα και από το 9.236, συμπεραίνουμε ότι το p-value είναι μεγαλύτερο του 0.1
 - Επομένως δεν απορρίπτουμε την μηδενική υπόθεση
- Κάνοντας χρήση του Minitab, προκύπτει το ίδιο αποτέλεσμα, εφόσον το αντίστοιχο p-value είναι ίσο με 0.673, δηλαδή πολύ μεγαλύτερο του 0.05, επομένως δεν απορρίπτουμε την H_0

		Test		Contribution	
Category	Observed	Proportion	Expected	to Chi-Sq	
1	341	0.36	360	1.00278	
2	255	0.26	260	0.09615	
3	153	0.15	150	0.06000	
4	77	0.07	70	0.70000	
5	96	0.09	90	0.40000	
6	78	0.07	70	0.91429	
N	DF	Chi-Sq	P-Value		
1000	5	3.17322	0.673		

Συμπεραίνουμε ότι οι διαφορές της δημοσκόπησης από τα αποτελέσματα των εκλογών δεν είναι στατιστικά σημαντικές σε επίπεδο εμπιστοσύνης 95%

12.4.4. Έλεγχος πίνακα

Ο έλεγχος είναι παρόμοιος με τον έλεγχο προσαρμογής

Ας τον εφαρμόσουμε με ένα παράδειγμα ([Chase & Dummer, 1992](#)) που παρουσιάζει τον αριθμό των μαθητών που θεωρούν τον αντίστοιχο στόχο ως πιο σημαντικό σε τρεις διαφορετικές τάξεις

Στόχοι	Τάξη 4	Τάξη 5	Τάξη 6
Ακαδημαϊκή επίδοση	49	50	69
Δημοφιλία	24	36	38
Αθλητική επίδοση	19	22	28

Ο πίνακας αυτός παρουσιάζει τις παρατηρούμενες συχνότητες (Observed ή O)

Τι θα μπορούσαμε να κάνουμε για να δούμε εάν υπάρχουν διαφορές στο πως βλέπουν τους στόχους οι διαφορετικές τάξεις;

Αρχικά, ας υπολογίσουμε τα αθροίσματα στηλών και γραμμών

Στόχοι	Τάξη 4	Τάξη 5	Τάξη 6	Σύνολο
Ακαδημαϊκή επίδοση	49	50	69	168
Δημοφιλία	24	36	38	98
Αθλητική επίδοση	19	22	28	69
Σύνολο	92	108	135	335

Τώρα μπορούμε να υπολογίσουμε τα ποσοστά σε κάθε τάξη (βήμα που δεν είναι απαραίτητο), για να δούμε εάν μας λένε τίποτα χρήσιμο

Στόχοι	Τάξη 4 (%)	Τάξη 5 (%)	Τάξη 6 (%)	Σύνολο (%)
Ακαδημαϊκή επίδοση	53.3	46.3	51.1	50.1
Δημοφιλία	26.1	33.3	28.1	29.3
Αθλητική επίδοση	20.7	20.4	20.7	20.6
Σύνολο	100	100	100	100

Φαίνεται να υπάρχουν διαφορές λοιπόν, το ερώτημα όμως είναι εάν οι διαφορές αυτές είναι στατιστικά σημαντικές

Επόμενο βήμα είναι να υπολογίσουμε τις αναμενόμενες (Expected ή E) συχνότητες (αριθμό μαθητών) σε κάθε κελί, εάν δεν υπήρχαν διαφορές ανάμεσα στις τάξεις, με βάση τον ακόλουθο μαθηματικό τύπο

Αναμενόμενη συχνότητα σε ένα κελί = Άθροισμα σειράς × Άθροισμα στήλης / Συνολικό άθροισμα

Στόχοι	Τάξη 4	Τάξη 5	Τάξη 6	Σύνολο
Ακαδημαϊκή επίδοση	46.1	54.2	67.7	168
Δημοφιλία	26.9	31.6	39.5	98
Αθλητική επίδοση	18.9	22.2	27.8	69
Σύνολο	92	108	135	335

Για παράδειγμα, η συχνότητα του πάνω αριστερά κελιού (Ακαδημαϊκή επίδοση στην Τάξη 4) βρέθηκε ως εξής

$$168 \times 92 / 335 = 46.1$$

Δεν μας πειράζει που οι αναμενόμενες συχνότητες δεν είναι ακέραιοι αριθμοί
 Τώρα μπορούμε να υπολογίσουμε την τιμή του χ^2 για τις διαφορές Observed
 μείον Expected

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

που ισούται με

$$\chi^2 = (49 - 46.1)^2/46.1 + (50 - 54.2)^2/54.2 + \dots + (28 - 27.8)^2/27.8 = 1.51$$

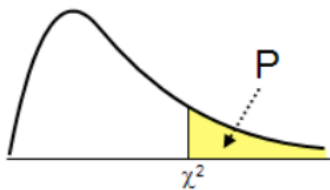
Για την εύρεση της κρίσιμης τιμής του χ^2 υπολογίζουμε τους βαθμούς ελευθερίας
 ως

$$df = (\text{Rows} - 1) \times (\text{Columns} - 1) = (3 - 1) \times (3 - 1) = 2 \times 2 = 4$$

χωρίς όμως να υπολογίζουμε τη γραμμή και τη στήλη των αθροισμάτων στον
 αριθμό γραμμών και στηλών

Σε επίπεδο εμπιστοσύνης 95%, η κρίσιμη τιμή της κατανομής χ^2 για 4 βαθμούς
 ελευθερίας ισούται με 9.488

Values of the Chi-squared distribution



	P										
DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697

Ως εκ τούτου δεν απορρίπτουμε τη μηδενική υπόθεση ουδεμίας διαφοράς ανάμεσα στις τάξεις και τους στόχους

Βιβλιογραφία

Chase, M.A., & Dummer, G.M. (1992). The role of sports as a social status determinant for children. *Research Quarterly for Exercise and Sport*, 63, 418-424.
<https://doi.org/10.1080/02701367.1992.10608764>

Weiss, N. (2012). *Elementary statistics* (8th ed). Addison Wesley, Boston, USA.