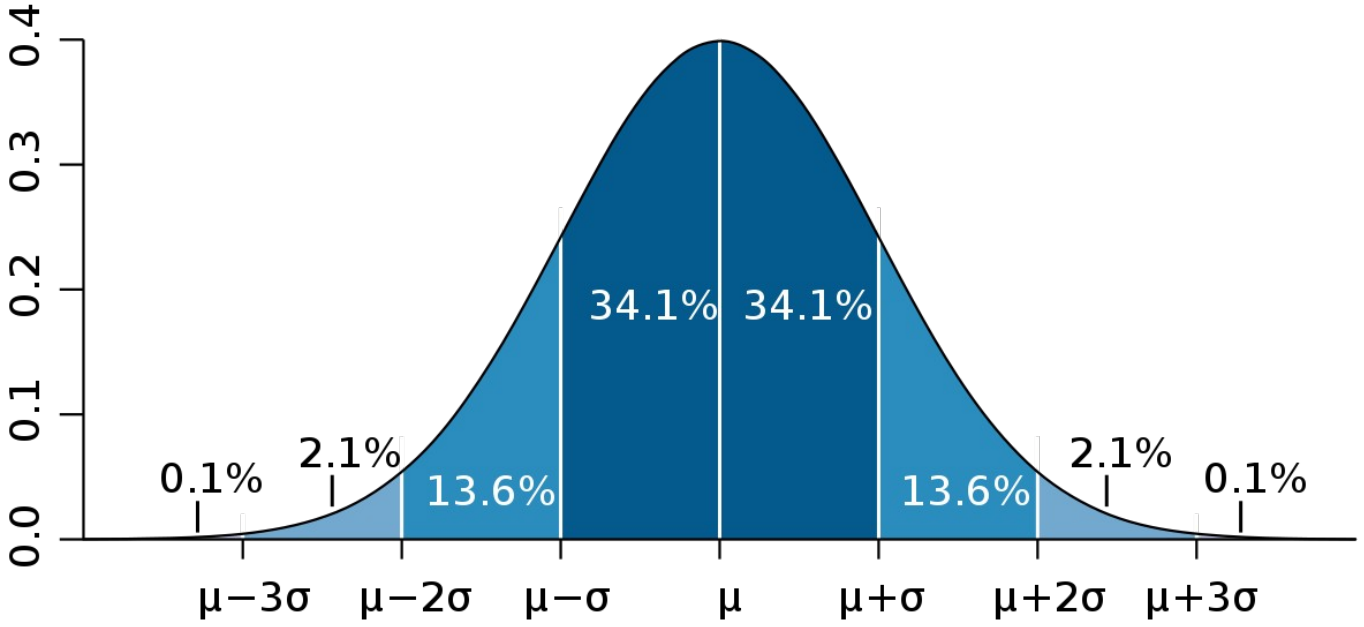
Θεωρία Πιθανοτήτων και Στατιστική



Διδάσκων: Επαμεινώνδας Διαμαντόπουλος
Επικοινωνία: epdiaman@ee.duth.gr

Περιεχόμενα 8ου μαθήματος

- Εισαγωγή στην Στατιστική
- Δειγματοληψία
- Περιγραφική Στατιστική
- Αμερόληπτοι και συνεπείς εκτιμητές.
- Εκτιμητές μέγιστης πιθανοφάνειας.
- Διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού
- Διάστημα εμπιστοσύνης για την αναλογία ενός χαρακτηριστικού στον πληθυσμό.
- Υπολογισμός μεγέθους δείγματος για μέση τιμή ή ποσοστό.

Γνωστικοί στόχοι 8ου μαθήματος

Στο τέλος αυτού του μαθήματος, ο φοιτητής πρέπει να είναι σε θέση :

- Να υπολογίζει τον εκτιμητή μέγιστης πιθανοφάνειας για μία άγνωστη παράμετρο.
- Να υπολογίζει διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού.
- Να υπολογίζει διάστημα εμπιστοσύνης για την αναλογία του πληθυσμού.
- Να μπορεί να υπολογίζει το απαραίτητο μέγεθος δείγματος για μία μελέτη.
- Να γνωρίζει τα βήματα μίας στατιστικής έρευνας και να αντιλαμβάνεται την ηθική σειρά των βημάτων που πρέπει να ακολουθήσει.

Στατιστική

Βασικές έννοιες

Στατιστικός πληθυσμός ή απλά **πληθυσμός** ονομάζεται κάθε σύνολο, τα στοιχεία του οποίου εξετάζουμε ως προς ένα ή περισσότερα χαρακτηριστικά τους. Τα στοιχεία του πληθυσμού ονομάζονται **μονάδες** ή **άτομα**. Ο πληθυσμός μπορεί να είναι **θεωρητικός** ή **πραγματικός.**

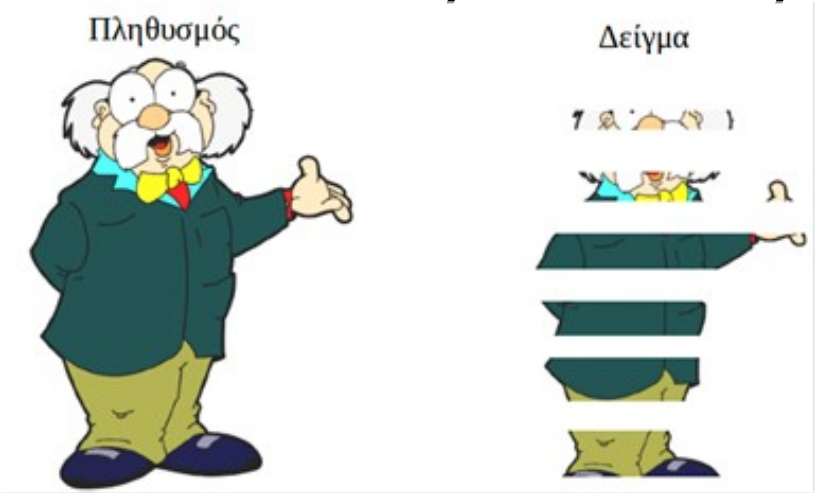
Δείγμα ονομάζεται το υποσύνολο του πληθυσμού το οποίο μπορούμε να καταγράψουμε υπό τους περιορισμούς (υλικούς και χρονικούς) της έρευνάς μας.

Οι κυριότερες μέθοδοι συλλογής στατιστικών δεδομένων είναι η **απογραφή** (census) και η **δειγματοληψία** (sampling).

Η **απογραφή** και η **δειγματοληψία**, μαζί με το **τυχαιοποιημένο πειραματικό σχέδιο** (randomized controlled trial) συνιστούν τις τρεις επιστημονικές μεθόδους με τις οποίες συλλέγονται στοιχεία και επεξεργάζονται με στατιστικές μεθόδους.

Όταν είναι εφικτή η απογραφή τότε αρκεί η **περιγραφική στατιστική** (descriptive statistics). Όταν η απογραφή είναι δύσκολη, οικονομικά και χρονικά ασύμφορη ή απλά αδύνατη, τότε είναι αναγκαία η επιλογή μιας μικρής ομάδας του πληθυσμού, δηλαδή ενός δείγματος. Συλλέγουμε τις παρατηρήσεις από το δείγμα και στη συνέχεια γενικεύουμε τα συμπεράσματα για ολόκληρο τον πληθυσμό με **επαγωγική** ή **συμπερασματική στατιστική** (inferential statistics).

Βασικές έννοιες



Το σφάλμα μίας δειγματοληψίας διαχωρίζεται σε **τυχαίο** και **συστηματικό**.

Τυχαίο σφάλμα δειγματοληψίας ονομάζεται η διαφορά μεταξύ των μετρήσεων του δείγματος και των πραγματικών μετρήσεων το οποίο θα υπάρχει στην έρευνά μας και δεν μπορούμε να το υπολογίζουμε επακριβώς εκτός αν καταφέρουμε να κάνουμε μία τέλεια εκτελεσμένη απογραφή!

Συστηματικό σφάλμα δειγματοληψίας ονομάζεται το σφάλμα που εμφανίζεται λόγω των σφαλμάτων που υπάρχουν στη σχεδίαση ή την υλοποίηση της δειγματοληψίας.

Στάδια δειγματοληψίας

Πλαίσιο δειγματοληψίας: ο φυσικός περιορισμός που ορίζεται στον πληθυσμό από το χρόνο και τόπο που διεξάγεται η δειγματοληψία.

Μέγεθος του δείγματος: ορίζεται είτε από το διαθέσιμο χρόνο και κόστος στην περίπτωση της μη πιθανοθεωρητικής δειγματοληψίας είτε με κατάλληλο υπολογισμό βάσει του επιθυμητού δειγματικού σφάλματος αν η δειγματοληψία πραγματοποιείται με κάποια πιθανοθεωρητική μέθοδο.

Πιθανοθεωρητική (propability sampling): η δειγματοληψία στην οποία κάθε μέλος του πληθυσμού έχει γνωστή πιθανότητα επιλογής πριν την υλοποίηση της δειγματοληψίας, δηλαδή είναι δυνατή η χρήση της Θεωρίας Πιθανοτήτων για τον υπολογισμό του τυχαίου σφάλματος της δειγματοληψίας.

Μη πιθανοθεωρητική (nonprobability sampling) ονομάζεται η δειγματοληψία στην οποία η πιθανότητα επιλογής των μελών του πληθυσμού είναι άγνωστη και δεν είναι δυνατή η εκ των προτέρων πιθανότητα επιλογής.

Είδη Πιθανοθεωρητικής Δειγματοληψίας

Απλή τυχαία δειγματοληψία (Simple Random Sampling): Κάθε μέλος του πληθυσμού έχει ίση πιθανότητα επιλογής στο δείγμα. Στην πράξη η απλή τυχαία δειγματοληψία συμβαίνει όταν υπάρχει η δυνατότητα να τοποθετηθεί ο πληθυσμός στη σειρά και μετά να επιλεχθεί το 10%- 15% με γεννήτρια τυχαίων αριθμών.

Συστηματική δειγματοληψία (Systematic Sampling): Η συστηματική δειγματοληψία συμβαίνει όταν θέλουμε να επιλέξουμε ένα τυχαίο δείγμα και είναι περισσότερο εύκολο να πάρουμε περιοδικό δείγμα αντί για τυχαίο, όπως για παράδειγμα σε δειγματοληψία μάρκετινγκ στην αγορά. Επιλέγεται μία αρχή με τυχαίο τρόπο και μετά επιλέγεται κάθε ν-οστό μέλος του καταλόγου. Στην πράξη: Τοποθετείται ο πληθυσμός στη σειρά 1, 2, ..., μετά επιλέγεται με τυχαίο τρόπο η πρώτη θέση (π.χ. 10), επιλέγεται το βήμα ανάλογα με το συνολικό μέγεθος του πληθυσμού (π.χ. 5) και μετά επιλέγεται το δείγμα από το $10^{ο}$, $15^{ο}$, $20^{ο}$ ... μέλος της σειράς.

Στρωματοποιημένη δειγματοληψία (Stratified Sampling): Ο ερευνητής ορίζει κάποια χαρακτηριστικά του πληθυσμού για τα οποία επιθυμεί οπωσδήποτε αναλογική εκπροσώπηση στο δείγμα του και επιλέγει απλό τυχαίο δείγμα αναλογικά από κάθε κατηγορία του πληθυσμού.

Είδη Μη Πιθανοθεωρητικής Δειγματοληψίας

Δειγματοληψία ευκολίας (convenience sampling): Το δείγμα αποτελείται από τις μονάδες του πληθυσμού που είναι διαθέσιμες εκείνη τη χρονική στιγμή.

Δειγματοληψία σκοπιμότητας (purposive sampling): Ένας εκπαιδευμένος δειγματολήπτης επιλέγει τις μονάδες του πληθυσμού που θεωρεί πως ανταποκρίνονται σε προκαθορισμένο προφίλ

Δειγματοληψία αναλογίας (quota sampling): Επιλογή του δείγματος έτσι ώστε να αντανακλάται σε αυτό η δημογραφική δομή του πληθυσμού ως προς ένα ή περισσότερα χαρακτηριστικά.

Δειγματοληψία χιονοστιβάδας (Snowball Sampling): Αρχική επιλογή ενός δείγματος με πιθανοθεωρητική μέθοδο και σε δεύτερο στάδιο συνέχιση της δειγματοληψίας από φίλο σε φίλο, από γείτονα σε γείτονα, κλπ. Συνιστάται στις περιπτώσεις που είναι επιθυμία του ερευνητή, το δείγμα να έχει κάποια συγκεκριμένα κοινωνικά ή πολιτικά χαρακτηριστικά.

Περιγραφική Στατιστική
(Descriptive Statistics)

Περιγραφή Δεδομένων

Το πρώτο μέλημα ενός ερευνητή είναι να περιγράψει με όσο το δυνατόν περισσότερη ακρίβεια, σαφήνεια και καθαρότητα τα δεδομένα τα οποία συνέλεξε. Ο τρόπος και οι μέθοδοι που θα χρησιμοποιηθούν για την περιγραφή αυτή εξαρτάται από το είδος των μεταβλητών. Συνοπτικά, στους παρακάτω πίνακες παρουσιάζονται τα βασικά μέτρα και γραφήματα που μπορούν να χρησιμοποιηθούν για την παρουσίαση των τιμών μίας μεταβλητής.

<table>
<tr><th>Είδος Μεταβλητής</th><th colspan="2">Προτεινόμενα Υπολογιστικά Μέτρα</th><th>Προτεινόμενα Γραφήματα</th></tr>
<tr><td rowspan="2">Ποιοτική
(όπως χρώμα ματιών, φύλο κ.α.)</td><td colspan="2" rowspan="2">Πίνακας Συχνοτήτων</td><td>Ραβδόγραμμα</td></tr>
<tr><td>Κυκλικό Διάγραμμα</td></tr>
<tr><td rowspan="8">Ποσοτική
(όπως ύψος, βάρος κ.α.)</td><td rowspan="3">Μέτρα θέσης</td><td>Επικρατούσα Τιμή</td><td rowspan="8">Ιστόγραμμα και Πολύγωνο Συχνοτήτων
(Για διακριτές ποσοτικές με «λίγες» τιμές είναι αποδεκτό επίσης το ραβδόγραμμα και το κυκλικό διάγραμμα)</td></tr>
<tr><td>Μέση Τιμή</td></tr>
<tr><td>Διάμεση Τιμή</td></tr>
<tr><td rowspan="5">Μέτρα διασποράς</td><td>Εύρος</td></tr>
<tr><td>Διακύμανση</td></tr>
<tr><td>Τυπική Απόκλιση</td></tr>
<tr><td>Απόλυτη Απόκλιση</td></tr>
</table>

Πίνακας Συχνοτήτων: Διακριτή τ.μ.

Ρωτήθηκαν 20 γυναίκες για το πλήθος των παιδιών που έχουν και έδωσαν τις παρακάτω αποκρίσεις : 0, 0, 1, 2, 0, 0, 1, 2, 1, 1, 1, 2, 2, 0, 4, 2, 3, 1, 0. (α) Να συμπληρωθεί ο πίνακας συχνοτήτων των παραπάνω παρατηρήσεων. (β) Να γίνει το ραβδόγραμμα συχνοτήτων και το κυκλικό διάγραμμα συχνοτήτων των τιμών.

Πλήθος (x_i)	Συχνότητα (n_i)	Σχετική Συχνότητα (f_i) $\frac{n_i}{N}$	Αθροιστική Συχνότητα (N_i)	Αθροιστική Σχετική Συχνότητα (F_i)
0	6	0,3	6	0,3
1	7	0,35	13	0,65
2	5	0,25	18	0,9
3	1	0,05	19	0,95
4	1	0,05	20	1
Σύνολο	20	1	—	—

Πίνακας Συχνοτήτων: Συνεχής τ.μ.

Οι 50 εργάτες ενός εργοστασίου έχουν τις παρακάτω ηλικίες: 21 43 50 25 55 30 28 40 31 51 18 47 52 34 47 32 27 41 35 54 30 48 36 43 38 33 27 39 41 43 32 22 46 52 29 32 34 34 42 36 35 28 57 56 20 38 27 27 40 35.

α) Να ομαδοποιήσετε τις ηλικίες στις κλάσεις: [18,28), [28,38), [38,48) και [48,58).

β) Να κατασκευάσετε πίνακα συχνοτήτων και σχετικών συχνοτήτων.

γ) Να κατασκευάσετε το ιστόγραμμα και το πολύγωνο συχνοτήτων.

Κλάση	Κέντρο (x_i)	n_i	f_i	N_i	F_i

Μέτρα Θέσης

Επικρατούσα τιμή

Τα κυριότερα μέτρα θέσης ή κεντρικής τάσης είναι η μέση τιμή, η διάμεση τιμή και η επικρατούσα τιμή.

Επικρατούσα τιμή (Mode)
Η επικρατούσα τιμή σε ένα σύνολο δεδομένων είναι απλά η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης. Όταν δύο οι περισσότερες τιμές συμπίπτουν στη συχνότητα τότε ονομάζονται όλες επικρατούσες τιμές.

Παράδειγμα
Η επικρατούσα τιμή του δείγματος 1, 3, 3, 4, 5, 5, 6, 2, 3, 4, 3, 1, 5 είναι η τιμή 3 με συχνότητα 4.

Διάμεση τιμή

Διάμεση τιμή (Median)

Η διάμεση τιμή (ή διάμεσος) σε ένα σύνολο δεδομένων είναι απλά η μεσαία παρατήρηση αν το πλήθος των στοιχείων είναι περιττό ενώ είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων αν το πλήθος των στοιχείων είναι άρτιο. Για να βρούμε τη διάμεσο κάνουμε τα εξής βήματα:

α) Ταξινομούμε τις παρατηρήσεις από τη μικρότερη στη μεγαλύτερη.

β) Η μεσαία παρατήρηση βρίσκεται στη θέση (n + 1) / 2.
Αν το (n + 1) / 2 είναι ακέραιος τότε η διάμεσος είναι η παρατήρηση που βρίσκεται στη θέση αυτή , ενώ αν είναι δεκαδικός τότε παίρνουμε το ημιάθροισμα των δύο παρατηρήσεων που βρίσκονται στις γειτονικές θέσεις.

Παράδειγμα

Η διάμεσος των 23, 4, 16 είναι το 16 ενώ η διάμεσος των παρατηρήσεων 23, 10, 13, 7 είναι το (10 + 13) / 2 = 11,5.

Μέση τιμή

Το νόημα της μέσης τιμής

Ως μέση τιμή ονομάζεται κάθε στατιστικό με το οποίο περιγράφεται το κέντρο των παρατηρήσεων. Ωστόσο, το «κέντρο» ενός συνόλου παρατηρήσεων που αποτελείται από πολλούς αριθμούς δεν ορίζεται μονοσήμαντα.

Ένας πρακτικός τρόπος ερμηνείας και κατανόησης της μέσης τιμής είναι η αναγνώρισή της ως το μέγεθος που θα μπορούσε να αντικαταστήσει το σύνολο των παρατηρήσεων ώστε να προκύπτει το ίδιο συνολικό αποτέλεσμα στο φυσικό πλαίσιο που ορίζεται η μεταβλητή.

Ο αριθμητικός μέσος αν και είναι η περισσότερο συχνή και απλή επιλογή δεν είναι πάντα η περισσότερο σωστή στην πράξη. Ανάλογα με το είδος των μονάδων μέτρησης της μεταβλητής, για τον υπολογισμό της μέσης τιμής μπορεί να επιλεχθεί:

- Ο αριθμητικός μέσος όταν η μεταβλητή εκφράζει καθαρές μονάδες (π.χ. ύψος ή βάρος)
- Ο αρμονικός μέσος όταν η μεταβλητή εκφράζει ρυθμό μεταβολής (π.χ. ταχύτητα)
- Ο γεωμετρικός μέσος όταν η μεταβλητή εκφράζει ποσοστιαίες μεταβολές (π.χ. επιτόκιο)

Αριθμητικός Μέσος

Αριθμητικός Μέσος (Mean ή Average)

Αν $x_1, x_2, \ldots, x_n$ είναι οι παρατηρήσεις του δείγματός μας τότε ο αριθμητικός μέσος ορίζεται να είναι:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

όπου n το μέγεθος του δείγματος ενώ αν τα στοιχεία $x_1, x_2, \ldots, x_n$ είναι όλος ο πληθυσμός για το οποίο γίνεται η έρευνα τότε γράφουμε:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Παράδειγμα

Αν 10 μαθητές έχουν βαθμολογία στη Στατιστική 12, 15, 10, 18, 17, 19, 15, 20, 13, 15 τότε η μέση βαθμολογία των μαθητών είναι:

$$\bar{x} = \frac{12 + 15 + 10 + 18 + 17 + 19 + 15 + 20 + 13 + 15}{10} = 15{,}4$$

Αριθμητικός Μέσος

3, 4, 5, 100

$\bar{x} = 4$ $\bar{x}_1 = \frac{112}{4} = 28$

Παρατήρηση
Ο αριθμητικός μέσος επηρεάζεται δυσανάλογα από τις πολύ μεγάλες ή τις πολύ μικρές παρατηρήσεις.

Πράγματι, η πρόσθεση ενός πολύ μεγάλου αριθμού στον αριθμητή του κλάσματος που ορίζει τη μέση τιμή θα τον αυξήσει δυσανάλογα σε σχέση με την αύξηση στον παρονομαστή η οποία θα είναι μόνο μία μονάδα.

Στην πράξη, αν υπάρχουν ιδιάζουσες τιμές στο δείγμα μας (πολύ μεγάλες ή πολύ μικρές παρατηρήσεις) τότε η μέση τιμή δεν αποτελεί αντιπροσωπευτικό στατιστικό του «κέντρου» των παρατηρήσεων και υπάρχουν οι εξής εναλλακτικές για την εκτίμηση του «κέντρου» της κατανομής:

(α) Ο υπολογισμός του αποκομμένου μέσου
(αφαιρείται το 5% έως 10% των πιο ακραίων παρατηρήσεων)

(β) η χρήση της διαμέσου (median)

Αρμονικός Μέσος

Θ → Ξ 60 Km/h
Ξ → Θ 40 Km/h

Αρμονικός Μέσος (Harmonic Mean)

Ο αρμονικός μέσος των παρατηρήσεων $x_1, x_2, \ldots, x_\nu$ ορίζεται να είναι η ποσότητα

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

1 Km $\frac{1}{60}$, 1 Km $\frac{1}{40}$

Ο αρμονικός μέσος είναι το κατάλληλο στατιστικό για τον υπολογισμό της μέσης τιμής όταν οι παρατηρήσεις εκφράζουν ρυθμούς μεταβολής.

Παράδειγμα

Ένα όχημα ταξιδεύει μία συγκεκριμένη διαδρομή με ταχύτητα 60 km/h (σε χρόνο t_1) και μετά επαναλαμβάνει την ίδια διαδρομή με 40 km/h (σε χρόνο t_2). Η μέση του ταχύτητα είναι

$$\bar{x}_h = \frac{2}{\frac{1}{60} + \frac{1}{40}} = 48 \text{ km/h}$$

δηλαδή αν ταξιδέψει με 48km/h θα καλύψει την απόσταση των δύο διαδρομών στον ίδιο χρόνο ($t_1 + t_2$). Είναι αξιοσημείωτο πως η τιμή αυτή διαφέρει από τον αριθμητικό μέσο (50 km/h).

Γεωμετρικός Μέσος

Γεωμετρικός Μέσος (Geometric Mean)

Ο γεωμετρικός μέσος των παρατηρήσεων $x_1, x_2, \ldots, x_n$ ορίζεται να είναι η ποσότητα

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Ο γεωμετρικός μέσος χρησιμοποιείται για τον υπολογισμό της μέσης τιμής ποσοστιαίων μεταβολών.

Παράδειγμα

Ο γεωμετρικός μέσος των αριθμών 3 και 5 είναι: $\bar{x}_g = \sqrt{3 \cdot 5} \approx 3{,}9$

ενώ ο γεωμετρικός μέσος των αριθμών 3, 5 και 8 είναι $\bar{x}_g = \sqrt[3]{3 \cdot 5 \cdot 8} \approx 4{,}9$

Άσκηση

Μία μετοχή που αξίζει 100 ευρώ κερδίζει τον πρώτο χρόνο 10%, το δεύτερο χρόνο 15% και τον τρίτο χρόνο 20%. Να βρεθεί η μέση ποσοστιαία αύξηση της μετοχής.

Διάμεση Τιμή και Αριθμητικός Μέσος

Η διαφοροποίηση της διάμεσου με τη μέση τιμή ως ένδειξη ασυμμετρίας
Σε μία πλήρως συμμετρική κατανομή η μέση τιμή και η διάμεση τιμή πρέπει να ταυτίζονται.

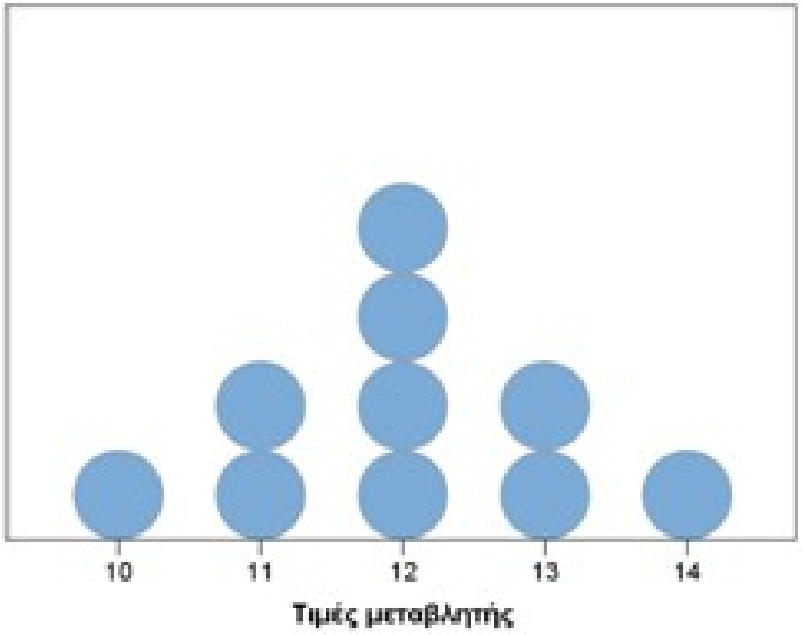
Το αντίθετο δεν ισχύει.
Μπορεί μία κατανομή να έχει μέση τιμή ίση με τη διάμεσο αλλά να είναι ασύμμετρη. Π.χ. αυτό συμβαίνει με τις παρατηρήσεις -2, -1, 0, 0, 3

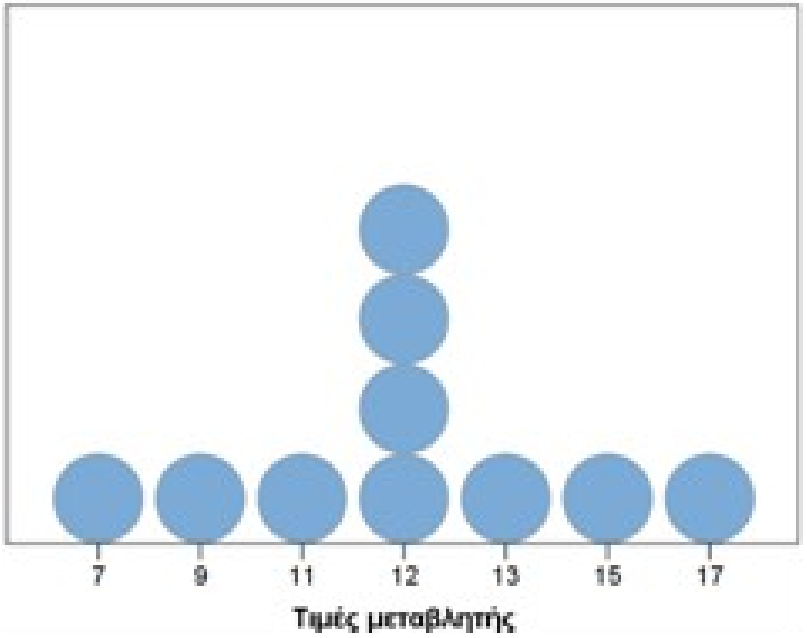
Στην πράξη, ωστόσο δεν συμβαίνει συχνά να είναι ίσες η μέση τιμή με τη διάμεσο. Αν η διάμεση τιμή είναι μικρότερη από τη μέση τιμή τότε αυτό μπορεί να ερμηνευθεί ως ένδειξη θετικής συμμετρίας (ουρά προς τα δεξιά της κατανομής) ενώ αν η διάμεση τιμή είναι μεγαλύτερη από τη μέση τιμή αυτό αποτελεί ένδειξη αρνητικής συμμετρίας (ουρά προς τα αριστερά της κατανομής).

Μέτρα Διασποράς

Μέτρα Διασποράς

Η περιγραφή της διασποράς μίας ομάδας παρατηρήσεων είναι απαραίτητη καθώς η μέση τιμή δεν δίνει πλήρη εικόνα για τη φύση της κατανομής. Χαρακτηριστικά, στην επόμενη εικόνα παρουσιάζονται δύο δείγματα με 10 τιμές που έχουν το ίδιο κέντρο (12) αλλά διαφορετική διασπορά.





Μέτρα Διασποράς

Με τη γενική ονομασία “Μέτρα Διασποράς” περιγράφουμε όλα τα στατιστικά που αποσκοπούν στην περιγραφή της διασποράς των παρατηρήσεων. Τα κυριότερα είναι τα εξής:

- Εύρος R (Range)
- Ενδοτεταρτημοριακό Εύρος IR (Interquartile Range)
- Μέση απόκλιση MAD (Mean Absolute Deviation)
- Διακύμανση Var (Variance)
- Τυπική απόκλιση StDev (Standard Deviation)

Εύρος

Το εύρος ενός δείγματος είναι απλά η διαφορά της μέγιστης από την ελάχιστη τιμή του.

R = Max – Min.

Το εύρος συνήθως συμβολίζεται με R από την αγγλική λέξη Range.

Παράδειγμα

Το εύρος των παρατηρήσεων -2, 0, 5, 4, 9, 11 είναι 11 – (-2) = 13.

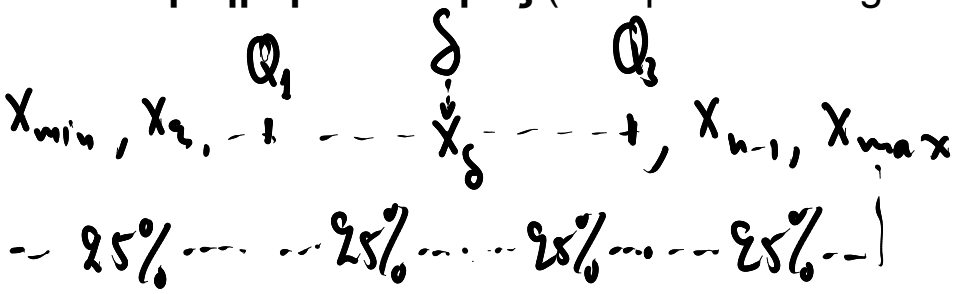
Ενδοτεταρτημοριακό Εύρος

Δύο σημαντικά σημεία σε μία κατανομή είναι αυτά που βρίσκονται στο 25ο και στο 75ο σημείο της κατανομής, τα οποία ονομάζονται **πρώτο** και **τρίτο τεταρτημόριο** και συμβολίζονται Q_1 και Q_3 αντίστοιχα. Το **πρώτο τεταρτημόριο (quartile) Q_1** βρίσκεται στην **(n + 1)/4 θέση** των ταξινομημένων παρατηρήσεων και αν ο αριθμός αυτός δεν είναι ακέραιος τότε υπολογίζουμε το ημιάθροισμα των στοιχείων που βρίσκονται στις δύο γειτονικές θέσεις. Ανάλογα, το **τρίτο τεταρτημόριο Q_3** βρίσκεται στην **3(n + 1)/4 θέση** και υπολογίζεται με τον ίδιο τρόπο.

Από τον ορισμό των Q_1, Q_3 είναι φανερό πως:

Μεταξύ του πρώτου και του τρίτου τεταρτημορίου βρίσκονται οι μισές παρατηρήσεις.

Ως **ενδοτεταρτημοριακό εύρος** (Intequartile Range – IR) ορίζεται η διαφορά **$IR = Q_3 - Q_1$**.



Ενδοτεταρτημοριακό Εύρος

Σημείωση
Ανάλογα με τα τεταρτημόρια ορίζονται:

(α) Τα πεμπτημόρια ως οι παρατηρήσεις στην 20%, 40%, 60% και 80% θέση της κατανομής. Η θέση τους στα ταξινομημένα δεδομένα είναι η α(v+1)/5, α = 1, 2, 3, 4.

(β) Τα δεκατημόρια ως οι παρατηρήσεις που βρίσκονται στην 10%, 20%, …, 90% θέση της κατανομής. Η θέση τους στα ταξινομημένα δεδομένα είναι η α(v+1)/10, α = 1, .., 9.

(γ) Τα εκατοστημόρια (percentiles) ως οι παρατηρήσεις που βρίσκονται στην 1%, 2%, …, 99% θέση της κατανομής. Η θέση τους στα ταξινομημένα δεδομένα είναι η α(v + 1)/100, α =1, …, 99.

Θηκόγραμμα

Το **θηκόγραμμα** (box plot) είναι ένα απλό γράφημα που δημιουργείται από τους 5 αριθμούς:

- Μέγιστη τιμή
- Τρίτο τεταρτημόριο (Q_3)
- Διάμεσος
- Πρώτο τεταρτημόριο (Q_1)
- Ελάχιστη τιμή.

Αποτελείται από ένα ορθογώνιο που ξεκινά από το Q_1 και τελειώνει στο Q_3, άρα έχει ύψος ίσο με το ενδοτεταρτημοριακό εύρος Q_3 - Q_1, το οποίο δείχνει το εύρος μέσα στο οποίο βρίσκονται οι μισές από τις παρατηρήσεις.

Η διάμεσος σχεδιάζεται ως μία έντονη γραμμή ενώ επιπλέον, έχει δύο ευθύγραμμα τμήματα που ξεκινούν από το μέσο της πλευράς του ορθογωνίου και τελειώνουν στη μέγιστη και στην ελάχιστη τιμή.

Ένας προσεκτικός αναγνώστης με την παρατήρηση του θηκογράμματος μπορεί να αντιληφθεί τη μορφή της κατανομής των τιμών.

Θηκόγραμμα

Μία παρατήρηση ονομάζεται **ιδιάζουσα τιμή (outlier)** όταν η απόστασή της από το πιο κοντινό τεταρτημόριο είναι μεγαλύτερη από το 1,5 · IR (ενδοτεταρτημοριακό εύρος).

Μία παρατήρηση ονομάζεται **ακραία τιμή** (extreme value) όταν η απόστασή της από το πιο κοντινό τεταρτημόριο είναι μεγαλύτερη από το 3 · IR.

Ένα θηκόγραμμα μπορεί να χρησιμοποιηθεί για την ανίχνευση ιδιαζόντων και ακραίων παρατηρήσεων.

Άσκηση

1. (α) Να βρεθεί το πρώτο, το τρίτο τεταρτημόριο και το ενδοτεταρτημοριακό εύρος των παρατηρήσεων 0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 7, 2.

(β) Να γίνει το θηκόγραμμα των τιμών.

Κώδικας R

```
x = c(0, 0, 1, 1, 1, 2, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 2, 7, 2)
summary(x)
boxplot(x, col = c("olivedrab"), horizontal = TRUE, names = c("Βαθμολογίες"))
```

Μέση απόκλιση, διακύμανση και τυπική απόκλιση

Μέση (απόλυτη) απόκλιση (mean absolute deviation) των παρατηρήσεων ονομάζεται η ποσότητα

$$\text{MAD} = \frac{1}{n}\sum_{k=1}^{n} |x_k - \bar{x}|$$

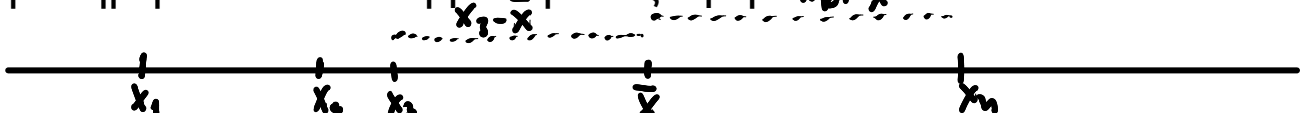
Διακύμανση ή διασπορά των παρατηρήσεων ονομάζεται η ποσότητα,

$$s^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})^2$$

Τυπική απόκλιση ονομάζεται η τετραγωνική ρίζα της διακύμανσης ή

$$s = \sqrt{s^2}$$

Κοινός στόχος και των τριών αυτών στατιστικών είναι να περιγράψουν την απόσταση των παρατηρήσεων από τη μέση τους τιμή.



Διακύμανση και τυπική απόκλιση

Άσκηση 2

Να βρεθεί η διακύμανση και η τυπική απόκλιση των 20 παρατηρήσεων

1, 0, 1, 1, 1, 5, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 4, 2 & 3.

Κώδικας R

```
x = c(1, 0, 1, 1, 1, 5, 3, 4, 1, 2, 2, 3, 3, 2, 1, 0, 1, 4, 2, 3)
summary(x)
sum((x - mean(x))^2)/length(x)
```

$$s^2 = \frac{1}{20}\sum_{i=1}^{20}(x_i-\bar{x})^2$$

$$\bar{x} = \frac{1+0+\cdots+3}{20} = \frac{40}{20} = 2$$

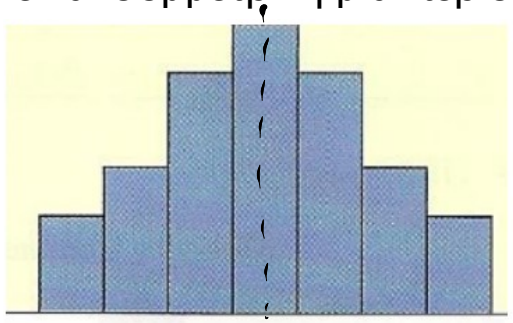
$$s^2 = \frac{1}{20}\cdot\left[2\cdot(0-2)^2 + 7\cdot(1-2)^2 + 4\cdot(2-2)^2 + 4(3-2)^2 + 2\cdot(4-2)^2 + 1\cdot(5-2)^2\right] =$$

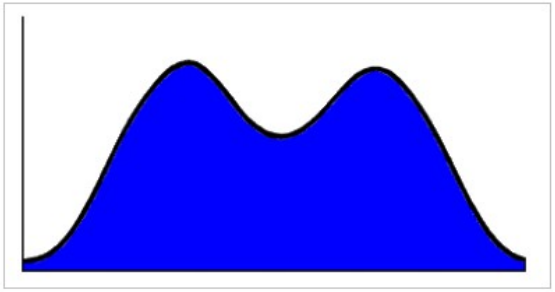
$$= \frac{1}{20}(8+7+0+4+8+9) = \frac{1}{20}\cdot 36 = 1{,}8. \quad \text{και} \quad s = \sqrt{1{,}8} = 1{,}34.$$

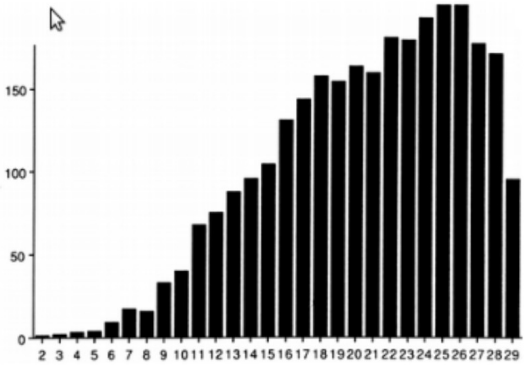
Ασυμμετρία και κυρτότητα κατανομής

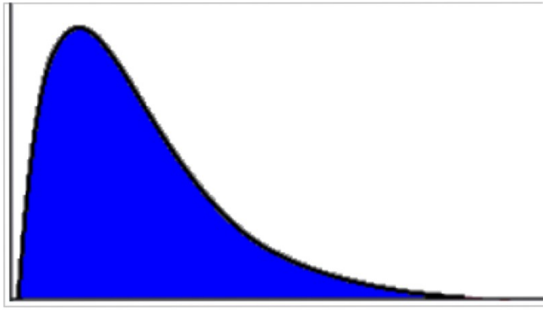
Ασυμμετρία μίας κατανομής (Skewness)

Μία κατανομή συχνοτήτων (ή σχετικών συχνοτήτων) ονομάζεται συμμετρική όταν είναι φανερό πως υπάρχει ένας κατακόρυφος άξονας ο οποίος λειτουργεί ως καθρέπτης της μισής κατανομής στην άλλη μισή. Χαρακτηριστικό παράδειγμα είναι η κανονική κατανομή χωρίς αυτό να σημαίνει πως δεν μπορεί να είναι συμμετρική μία περισσότερο “ανώμαλη” κατανομή.









Ασυμμετρία και κυρτότητα μίας κατανομής

Συντελεστής ασυμμετρίας (skew)

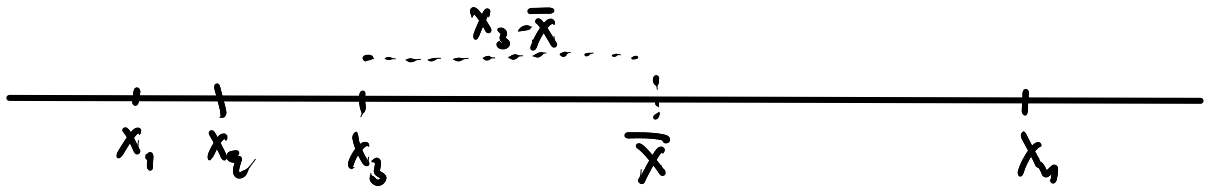
$$\gamma = \frac{\mu_3}{s^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

Συντελεστής κυρτότητας (kurtosis)

$$\alpha = \frac{\mu_4}{s^4} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

...όπου $\mu_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$, $k = 1, 2, \ldots$ οι κεντρικές ροπές k τάξης.

Στη βιβλιογραφία συναντώνται και οι εκδοχές με αποκλειστική χρήση των κεντρικών ροπών: $\gamma = \frac{\mu_3}{(\mu_2)^{3/2}}$, $\alpha = \frac{\mu_4}{\mu_2^2}$

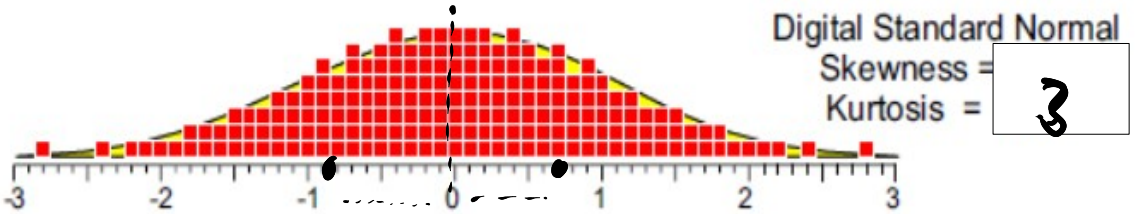


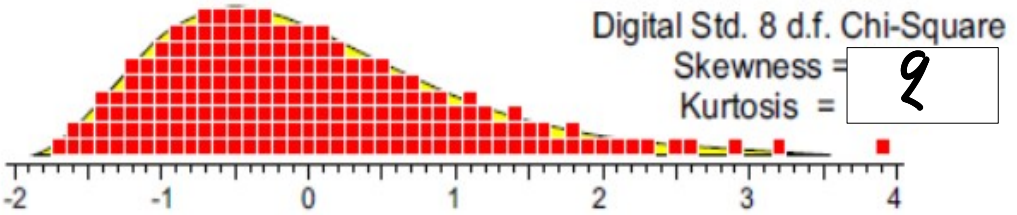
Άσκηση 1

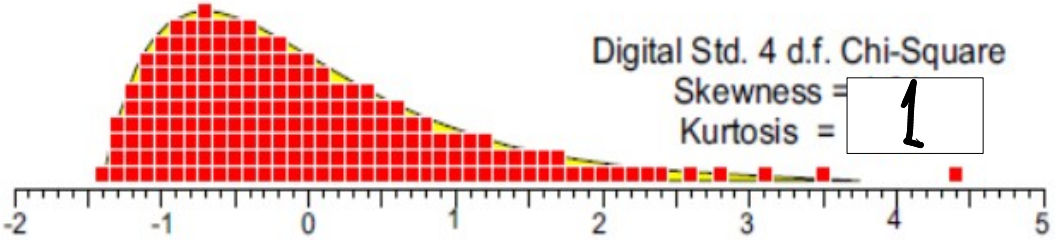
Αντιστοιχίστε τις παρακάτω τιμές ασυμμετρίας και κυρτότητας με τις 4 δειγματικές κατανομές του σχήματος:

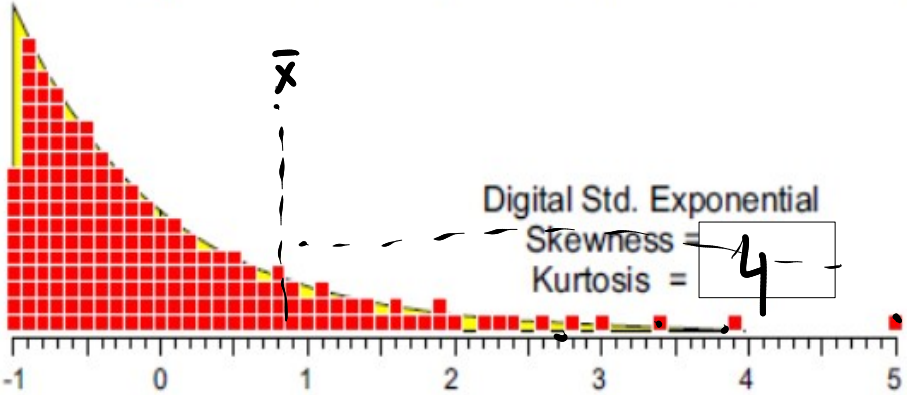
Ασυμμετρία γ	Κυρτότητα α	
1,31	5,18	1
0,93	4,03	2
0,00	2,88	3
1,84	7,34	4

$\sum (x_i - \bar{x})^3$ $\sum (x_i - \bar{x})^4$







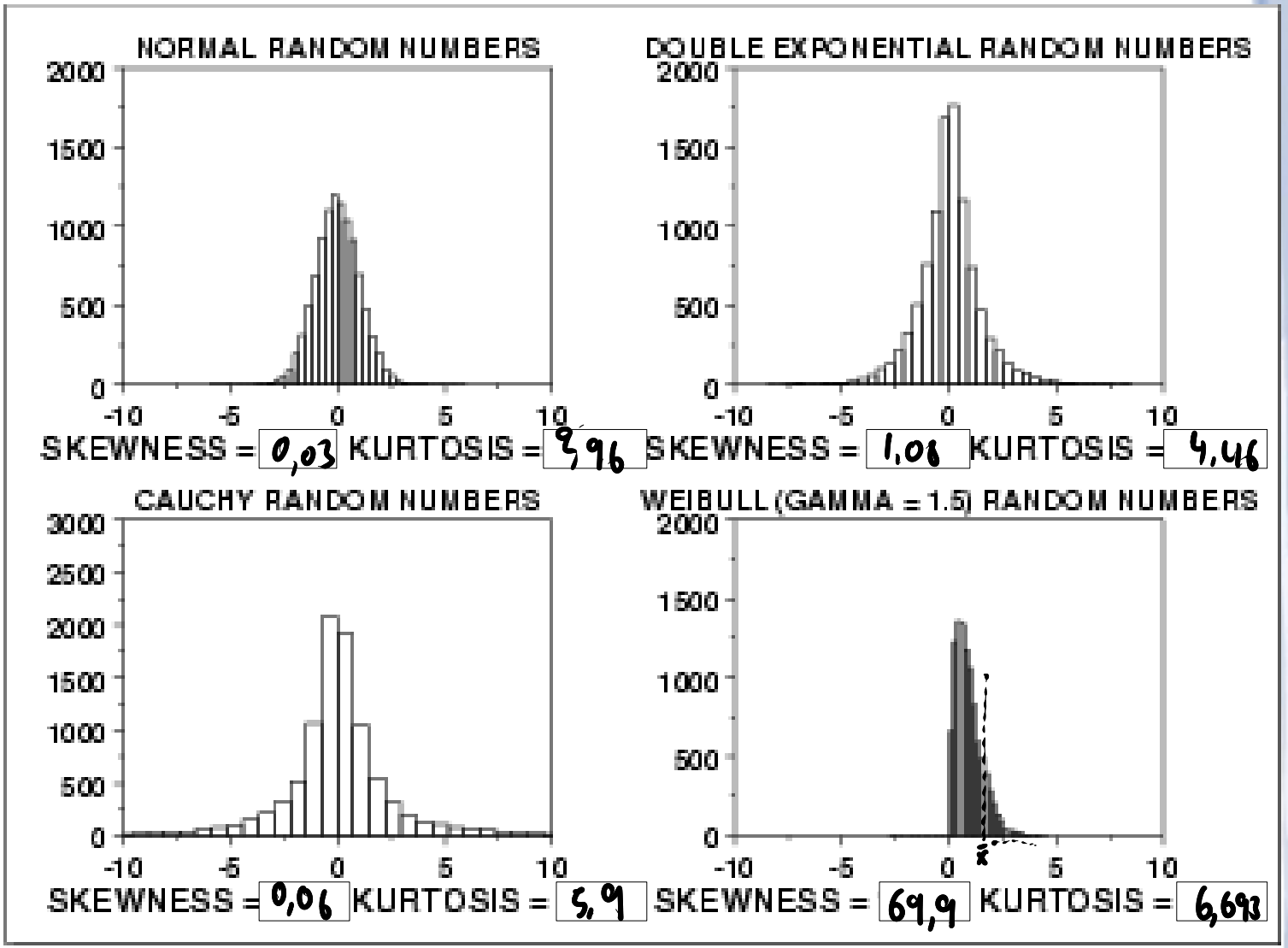


Άσκηση 2

Αντιστοιχίστε τις παρακάτω τιμές ασυμμετρίας και κυρτότητας με τις 4 δειγματικές κατανομές του σχήματος:

Ασυμμετρία γ	Κυρτότητα α
~~69,9~~	~~6.693~~
0,06	5,90
0,03 .	2,96 .
1,08	4,46 .

$\Sigma(x_i - \bar{x})^4$



Επαγωγική Στατιστική
(Inferential Statistics)

Από το δείγμα στον πληθυσμό

Όταν η πρόσβαση στο σύνολο του πληθυσμού είναι μία εφικτή επιλογή τότε η περιγραφική στατιστική είναι αρκετή για έναν ερευνητή: με τη χρήση κατάλληλων διαγραμμάτων και πινάκων ο ερευνητής αποκτά πλήρη και συγκεκριμένη εικόνα για τα στοιχεία του πληθυσμού που έχει συλλέξει και μπορεί να πάρει τις επιχειρηματικές (ή άλλες) αποφάσεις του.

Ωστόσο, στις περισσότερες πραγματικές περιπτώσεις συλλογής δεδομένων δεν είναι δυνατή η κάλυψη όλου του πληθυσμού λόγω χρονικών, τεχνικών και οικονομικών περιορισμών και επιλέγεται η κάλυψη ενός υποσυνόλου του πληθυσμού, δηλαδή η μέτρηση ενός δείγματος.

Ανακύπτει με φυσικό τρόπο η ανάγκη της γενίκευσης των παρατηρήσεων από το δείγμα σε όλον τον πληθυσμό με τρόπο ώστε να προσδιορίζεται το σφάλμα που φανερά αντιστοιχεί σε αυτήν τη γενίκευση.

Στο σημείο αυτό, εμφανίζεται η **επαγωγική στατιστική** και οι μέθοδοι της.

Από το δείγμα στον πληθυσμό

Με τον όρο **Επαγωγική Στατιστική** (Inferential Statistics) περιγράφονται όλες οι στατιστικές διαδικασίες που έχουν ως σκοπό την εκμαίευση χρήσιμων δεδομένων για τον πληθυσμό από τα στοιχεία ενός αντιπροσωπευτικού δείγματός του.

Χρησιμοποιούμε Επαγωγική Στατιστική όταν δεν έχουμε πρόσβαση στον πληθυσμό.

Απαραίτητη προϋπόθεση είναι η επιλογή ενός δείγματος που να αντιπροσωπεύει όσο το δυνατόν καλύτερα τον πληθυσμό που μας ενδιαφέρει.

Αν ο ερευνητής δεν είναι βέβαιος πως το δείγμα του αντιπροσωπεύει τον πληθυσμό ως προς τα χαρακτηριστικά που τον ενδιαφέρουν τότε το αποτέλεσμα θα είναι μακριά από την αλήθεια και θα οδηγήσει σε λανθασμένες αποφάσεις.

Όριο κατά πιθανότητα

Όριο κατά πιθανότητα

Έστω η ακολουθία των τ.μ. $X_n = \begin{cases} 1, \text{ με πιθανότητα } \frac{1}{n} \\ 0, \text{ με πιθανότητα } 1 - \frac{1}{n} \end{cases}$

Δραστηριότητα

Αναλογιστείτε ποιο μπορεί να είναι το όριο των X_n.

$\lim_{n \to \infty} X_n = X$ με $\begin{matrix} 1, P(X=1)=0 \\ 0, P(X=0)=1 \end{matrix}$.

$P(|X_n - X|)$

Όριο κατά πιθανότητα

Λέμε ότι $X_n \to X$ κατά πιθανότητα όταν, καθώς $n \to \infty$, γίνεται όλο και πιο απίθανο να οι τ.μ. X_n να είναι μακριά από τη X.

Γράφουμε:

$$\text{plim}_{n\to\infty} X_n = X \leftrightarrow \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0, \forall \varepsilon > 0$$

Ιδιότητες

- $\text{plim}(\alpha X_n + b) = \alpha \cdot \text{plim}(X_n) + b$
- Αν $X_n \to X$ κατά πιθανότητα και g συνεχής, τότε $\text{plim}\, g(X_n) = g(X)$
- Αν $X_n \to X$, $Y_n \to Y$ κατά πιθανότητα και g συνεχής, τότε $\text{plim}\, g(X_n, Y_n) = g(X, Y)$

Δύο Χρήσιμες Ανισότητες

Ανισότητα Markov

Έστω ότι X είναι μία τ.μ. με θετικές τιμές. Τότε για κάθε α > 0, ισχύει

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}.$$

Απόδειξη (για X συνεχής τ.μ.)

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} x f(x)\,dx \\
&= \int_{0}^{a} x f(x)\,dx + \int_{a}^{\infty} x f(x)\,dx \\
&\geq \int_{a}^{\infty} x f(x)\,dx \geq \int_{a}^{\infty} a f(x)\,dx = a\int_{a}^{\infty} f(x)\,dx = a\Pr(X \geq a)
\end{aligned}$$

$$\Pr(X \geq a) \leq \mathrm{E}(X)/a$$

Πηγή: https://en.wikipedia.org/wiki/Markov%27s_inequality

Ανισότητα Markov

Έστω ότι X είναι μία τ.μ. με θετικές τιμές. Τότε για κάθε α > 0, ισχύει

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}.$$

Απόδειξη (για X διακριτή τ.μ.)

$$\begin{aligned} E(X) &= \sum_{x \geq a} xP(X = x) + \sum_{x < a} xP(X = x) \\ &\geq \sum_{x \geq a} aP(X = x) + 0 \\ &= a \sum_{x \geq a} P(X = x) \\ &= aP(X \geq a) \end{aligned}$$

Πηγή: https://math.dartmouth.edu/~m20x18/markov

Ανισότητα Markov

Άσκηση: Έστω η τμ $X \sim Exp(\lambda)$.
(α) Να εφαρμοστεί η ανισότητα Markov για να βρεθεί άνω φράγμα για την $P(X \geq \alpha)$, $\alpha > 0$.
(β) Να συγκριθεί η εκτίμηση με την πραγματική τιμή της $P(X \geq \alpha)$.

Λύση

$$X \sim Exp(\lambda), \quad E(X) = \frac{1}{\lambda}, \quad f(x) = \lambda e^{-\lambda x}.$$

(α) Markov: $$P(X \geq a) \leq \frac{E(X)}{a} = \frac{1}{\lambda a}.$$

(β) $$P(X \geq a) = \int_a^{+\infty} f(x)\,dx = \int_a^{+\infty} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_{x=a}^{x=+\infty} = e^{-\lambda a} \leq \frac{1}{\lambda a}$$

Ανισότητα Markov

Άσκηση: Έστω ότι ένα πείραμα έχει πιθανότητα επιτυχίας p και επαναλαμβάνεται μέχρι να συμβεί η πρώτη επιτυχία και η τμ X μετράει το πλήθος δοκιμών μέχρι την πρώτη επιτυχία.
(α) Να εφαρμοστεί η ανισότητα Markov για να βρεθεί άνω φράγμα για την $P(X \geq \alpha)$, $\alpha > 0$.
(β) Να συγκριθεί η εκτίμηση με την πραγματική τιμή της $P(X \geq \alpha)$.

Λύση
Υπόδειξη: $X \sim G_T(p)$ και $EX = 1/p$.

$$X \sim G_T(p), \quad E(X) = \frac{1}{p}, \quad P(X = a) = (1-p)^{a-1} \cdot p$$

$$(1-p)^{a-1} \leq \frac{1}{pa}.$$

(a) $$P(X \geq a) \leq \frac{E(X)}{a} = \frac{1}{pa}$$

(β) $$P(X \geq a) = \sum_{n=a}^{+\infty} P(X = n) = \sum_{n=a}^{+\infty} (1-p)^{n-1} \cdot p \overset{n-a=k}{=} p \sum_{k=0}^{+\infty} (1-p)^{k+a-1} = p \cdot (1-p)^{a-1} \frac{1}{p} = (1-p)^{a-1}.$$

Ανισότητα Chebyshev

Έστω ότι X είναι μία ΤΜ τέτοια ώστε $\sigma^2 < +\infty$, $\sigma^2 \neq 0$. Τότε για κάθε $k > 0$, ισχύει

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \text{ ή ισοδύναμα } P(|X - \mu| \geq k) \leq \frac{Var(X)}{k^2}.$$

Απόδειξη

Η απόδειξη της ανισότητας Chebyshev είναι μία κατάλληλη εφαρμογή της ανισότητας Markov:

$$P(|X - \mu| \geq k\sigma) = P((X - \mu)^2 \geq k^2\sigma^2)$$

$$\leq \frac{E(X - \mu)^2}{k^2\sigma^2}$$

$$= \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Ανισότητα Chebyshev

Έστω ότι X είναι μία τ.μ. τέτοια ώστε $\sigma^2 < +\infty$, $\sigma^2 \neq 0$. Τότε για κάθε $k > 0$, ισχύει

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Παραδείγματα:

- Αν k = 1, τότε η ανισότητα που προκύπτει είναι προφανής: $P(|X - \mu| \geq \sigma) \leq 1$. προφανής.
- Αν k = $\sqrt{2}$, τότε $P(|X - \mu| \geq \sqrt{2}\sigma) \leq \frac{1}{2}$, δηλαδή η πιθανότητα μία τιμή της X να βρίσκεται έξω από το διάστημα $(\mu - \sqrt{2}\sigma, \mu + \sqrt{2}\sigma)$ είναι μικρότερη από 0,5 = 50%.
- Αν k = 5, τότε $P(|X - \mu| \geq 5\sigma) \leq \frac{1}{25}$, δηλαδή η πιθανότητα μία τιμή της X να βρίσκεται έξω από το διάστημα $(\mu - 5\sigma, \mu + 5\sigma)$ είναι μικρότερη από 0,04 = 4%.

Ανισότητα Chebyshev

Άσκηση: Έστω η τμ $X \sim Exp(\lambda)$. Να εφαρμοστεί η ανισότητα Chebyshev για να βρεθεί άνω φράγμα για την $P(|X - 1/\lambda| \geq \beta)$, $\beta > 0$.

Λύση

$$X \sim Exp(\lambda), \quad E(X) = \mu = \frac{1}{\lambda}, \quad Var(X) = \frac{1}{\lambda^2}$$

$$P\left(\left|X - \frac{1}{\lambda}\right| \geq \beta\right) = P(|X - \mu| \geq \beta) \leq \frac{Var(X)}{\beta^2} = \frac{1}{\lambda^2 \beta^2}.$$

$$X \geq \beta + \mu$$

$$X \leq -\beta + \mu$$

Ανισότητα Chebyshev

Άσκηση: Το πλήθος των πελατών σε ένα κατάστημα έχει μέση τιμή 100 και διακύμανση 225. Να βρεθεί ένα άνω όριο για την πιθανότητα να υπάρχουν περισσότεροι από 120 ή λιγότεροι από 80 πελάτες στο κατάστημα.

Λύση

$$E(X) = \mu = 100, \quad Var(X) = 225$$

$$P(X > 120 \text{ ή } X < 80) = P(|X - 100| \geq 20) \leq \frac{Var(X)}{20^2} = \frac{225}{400} = 0,5625$$

Ανισότητα Chebyshev

Άσκηση (Ασθενής Νόμος των Μεγάλων Αριθμών - Weak Law of Large Numbers)

Έστω ότι $X_1, X_2, \ldots, X_n$ είναι ΤΜ με ίδια κατανομή πιθανότητας, τέτοια ώστε $\sigma^2 < +\infty$, $\sigma^2 \neq 0$. Τότε, να δείξετε ότι $\text{plim}_{n \to +\infty} \bar{X}_n = \mu$ ή ισοδύναμα ότι για κάθε $\varepsilon > 0$, είναι:

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \to 0, \text{ καθώς } n \to \infty.$$

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{Var(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n \cdot \varepsilon^2} \xrightarrow{n \to +\infty} 0 \Rightarrow \underset{n \to +\infty}{plim}\, \bar{X}_n = \mu$$

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Εκτιμητές

Εκτιμητές

Όταν γνωρίζουμε την κατανομή που ακολουθεί μία τυχαία μεταβλητή και την αντίστοιχη συνάρτηση πιθανότητας f(x), τότε έχουμε δυνατότητα υπολογισμού όλων των γεωμετρικών χαρακτηριστικών της, όπως είναι η αναμενόμενη τιμή, η διακύμανση κλπ.

Στην πράξη όμως, τις περισσότερες φορές, αν και γνωρίζουμε το είδος της κατανομής, δεν γνωρίζουμε επακριβώς τη συνάρτηση πιθανότητας, αλλά ούτε και τις παραμέτρους του πληθυσμού.

Στην περίπτωση αυτή, συλλέγουμε ένα δείγμα τιμών από τον πληθυσμό και προσπαθούμε να εκτιμήσουμε από αυτό τις άγνωστες παραμέτρους που μας ενδιαφέρουν.

Οι μέθοδοι που έχουν τον παραπάνω στόχο ανήκουν στο γνωστικό αντικείμενο της **Επαγωγικής Στατιστικής.**

Εκτιμητές

Στην επαγωγική στατιστική, αξιοποιούμε ένα δείγμα του πληθυσμού $x = (x_1, \ldots, x_n)$, για να ορίσουμε ένα δειγματικό στατιστικό $\theta_n = \theta_n(x)$ και να εκτιμήσουμε μία άγνωστη παράμετρο θ του πληθυσμού.

Παραδείγματα

Για την παράμετρο $\theta = \mu = E(X)$ χρησιμοποιούμε τον $T_n = (x_1 + x_2 + \ldots + x_n) / n$.

Για την παράμετρο $\theta = \sigma^2 = Var(X)$ χρησιμοποιούμε τον $\sigma_n^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2$

Αν $X \sim U(\alpha, \beta)$, τότε για το $\theta = \beta$, χρησιμοποιούμε το $T_n = \max\{x_1, x_2, \ldots, x_n\}$

$E(X) = \frac{1}{\lambda}$

Αν $X \sim Exp(\lambda)$, τότε για το $\theta = \lambda$, χρησιμοποιούμε το $\lambda_n = \frac{n}{x_1 + x_2 + \ldots + x_n}$

5, 3, 10. $\frac{5+3+10}{3} = 6$, $\lambda = \frac{3}{5+3+10}$

Ο εκτιμητής ως μία Τυχαία Μεταβλητή

Στην πράξη, ένα δείγμα του πληθυσμού $x = (x_1, \ldots, x_n)$, είναι ένα αριθμητικό διάνυσμα. Ωστόσο, στη θέση κάθε μίας παρατήρησης θα μπορούσε να είναι οποιαδήποτε άλλη από τον θεωρητικό πληθυσμό. Δηλαδή, το τυχαίο σύνολο τιμών x μπορεί να γίνει αντιληπτό ως ένα διάνυσμα τυχαίων μεταβλητών:

$$X = (X_1, \ldots, X_n).$$

Στο πλαίσιο αυτό, αναγνωρίζουμε και τον εκτιμητή θ_n ως μία τυχαία μεταβλητή η οποία αποδίδει εκτιμήσεις για την άγνωστη παράμετρο θ από το τυχαίο δείγμα που του παρέχεται.

Ο εκτιμητής ως μία Τυχαία Μεταβλητή

Όταν έχουμε έναν εκτιμητή θ_n, το βασικό ερώτημα δεν είναι:

«Ποιος είναι ο τύπος του;»

αλλά:

«Μπορώ να τον εμπιστευτώ;»

Δύο βασικά ερωτήματα:

α) Για σταθερό n, για κάθε δυνατή τιμή των X_i, ποια είναι η αναμενόμενη τιμή της κατανομής του $\theta_n = \theta_n(X_1, X_2, \ldots, X_n)$;

β) Καθώς το $n \to \infty$, με ποιον τρόπο το θ_n προσεγγίζει την άγνωστη παράμετρο μ του πληθυσμού;

Μεροληψία εκτιμητή

Έστω $\theta_n = \theta_n(X_1, X_2, \ldots, X_n)$ ένας εκτιμητής της παραμέτρου θ.

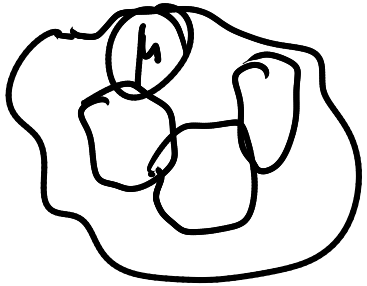
Ως **μεροληψία του θ_n** ορίζεται να είναι η ποσότητα:

$$\text{Bias}(\theta_n, \theta) = E_{x|\theta}(\theta_n) - \theta$$

$$= E_{x|\theta}(\theta_n - \theta),$$

Το $E_{x|\theta}$ σημαίνει ότι η αναμενόμενη τιμή υπολογίζεται δεδομένης συγκεκριμένης τιμής για την παράμετρο θ.

Αν **$\text{Bias}(\theta_n, \theta) = 0$**, ή ισοδύναμα **$E_{x|\theta}(\theta_n) = \theta$**, τότε ο εκτιμητής θ_n ονομάζεται **αμερόληπτος**.



$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Μία μέτρηση είναι αμερόληπτος εκτιμητής του μ

Αν μ είναι η άγνωστη μέση τιμή όλου του πληθυσμού και Χ μία μέτρηση ενός στοιχείου του πληθυσμού, τότε το Χ ακολουθεί την κατανομή του πληθυσμού και άρα

$$E(X) = \mu.$$

Συμπεραίνουμε:

$$\begin{aligned} Bias(X, \mu) &= E(X - \mu) \\ &= E(X) - \mu \\ &= \mu - \mu \\ &= 0. \end{aligned}$$

Δηλαδή, ο εκτιμητής που αντιστοιχεί σε μία απλή μέτρηση είναι αμερόληπτος.

Ο αρ. μέσος είναι αμερόληπτος εκτιμητής του μ

Αν μ είναι η άγνωστη μέση τιμή όλου του πληθυσμού και $X = (X_1, \ldots, X_n)$, τότε ο αριθμητικός μέσος $T_n = (X_1 + \ldots + X_n)/n$ είναι ένας αμερόληπτος εκτιμητής για τη μ. Πράγματι,

$$\begin{aligned}
\text{Bias}(T_n, \mu) &= E_{x|\mu}(T_n - \mu) \\
&= E_{x|\mu}(T_n) - \mu \\
&= E_{x|\mu}[(X_1 + \ldots + X_n)/n] - \mu \\
&= [E(X_1) + \ldots + E(X_n)]/n - \mu \\
&= (\mu + \ldots + \mu)/n - \mu \\
&= \mu - \mu \\
&= 0.
\end{aligned}$$

Συνεπείς εκτιμητές

Έστω θ_n, ένας εκτιμητής της παραμέτρου θ. Ο θ_n ονομάζεται **συνεπής** (consistent) όταν

$$\text{plim}_{n\to\infty}(\theta_n - \theta) = \lim_{n\to\infty} P(|\theta_n - \theta| > \varepsilon) = 0, \text{ για κάθε } \varepsilon > 0.$$

(η αύξηση του μεγέθους του δείγματος οδηγεί σε καλύτερη πρόβλεψη της παραμέτρου)

$\theta_n \xrightarrow{\text{plim}} \theta$

Ο αρ. μέσος είναι ένας συνεπής εκτιμητής του μ

Ο αριθμητικός μέσος T_n είναι ένας συνεπής εκτιμητής για τη μέση τιμή του πληθυσμού. Η απόδειξη γίνεται με τη βοήθεια του Κεντρικού Οριακού Θεωρήματος από το οποίο παίρνουμε $T_n \sim N(\mu, \sigma^2/n)$ ή $T_n - \mu \sim N(0, \sigma^2/n)$ και για κάθε $\varepsilon > 0$,

$$P(|T_n - \mu| > \varepsilon) = P(T_n - \mu > \varepsilon) + P(T_n - \mu < -\varepsilon)$$

$$= 2P(T_n - \mu > \varepsilon)$$

$$= 2[1 - P(T_n - \mu \leq \varepsilon)]$$

$$= 2[1 - P(Z < n^{1/2}\varepsilon/\sigma)]$$

$$= 2(1 - \Phi(n^{1/2}\varepsilon/\sigma)).$$

Από την τελευταία σχέση παίρνουμε:

$$\text{plim}_{n\to\infty} T_n = \lim_{n\to\infty} P(|T_n - \mu| > \varepsilon)$$

$$= \lim_{n\to\infty} 2(1 - \Phi(n^{1/2}\varepsilon/\sigma))$$

$$= 0.$$

Αμερόληπτος εκτιμητής για τη διακύμανση σ^2

Χρήσιμες ιδιότητες της διακύμανσης

Μία χρήσιμη ιδιότητα της διακύμανσης είναι η παρακάτω:

Θεώρημα
Αν X, Y ανεξάρτητες τότε Var(X + Y) = Var(X) + Var(Y)

Απόδειξη

$$
\begin{aligned}
\operatorname{Var}(X+Y) &= \mathrm{E}\left[(X+Y)^2\right] - (\mathrm{E}[X+Y])^2 \\
&= \mathrm{E}\left[X^2 + 2XY + Y^2\right] - (\mathrm{E}[X] + \mathrm{E}[Y])^2 \\
&= \mathrm{E}\left[X^2\right] + 2\,\mathrm{E}[XY] + \mathrm{E}\left[Y^2\right] - \left(\mathrm{E}[X]^2 + 2\,\mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}[Y]^2\right) \\
&= \mathrm{E}\left[X^2\right] + \mathrm{E}\left[Y^2\right] - \mathrm{E}[X]^2 - \mathrm{E}[Y]^2 \\
&= \operatorname{Var}(X) + \operatorname{Var}(Y).
\end{aligned}
$$

Χρήσιμες ιδιότητες της διακύμανσης

Πόρισμα 1
Αν $X_1, X_2, \ldots, X_n$ ανεξάρτητες τ.μ. με ίδια διακύμανση σ^2, τότε $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Απόδειξη

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Πόρισμα 2
Αν $X_1, X_2, \ldots, X_n$ ανεξάρτητες τ.μ. με ίδια διακύμανση σ^2, τότε $\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2$.

Απόδειξη

Προκύπτει άμεσα από την παρατήρηση πως $\mu_{\bar{X}} = \mu_X = \mu$ και το Πόρισμα 1.

Πηγή: https://en.wikipedia.org/wiki/Variance#Sum_of_uncorrelated_variables_(Bienaym%C3%A9_formula)

Ένας αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 2

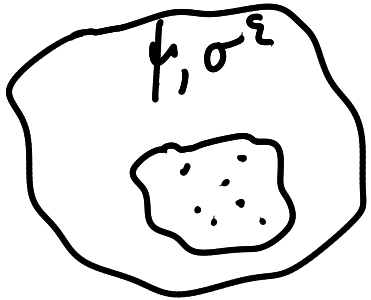
Αν η αναμενόμενη τιμή μ της ΤΜ X είναι γνωστή και σ^2 είναι η άγνωστη διακύμανση της, τότε η παράσταση

$$\sigma_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2$$

είναι αμερόληπτος εκτιμητής για το σ^2.

Απόδειξη

$$E(\sigma_n^2 - \sigma^2) = E(\sigma_n^2) - \sigma^2 = \frac{1}{n}\sum_{k=1}^{n} E(X_k - \mu)^2 - \sigma^2 = \frac{1}{n}\sum_{k=1}^{n} \sigma^2 - \sigma^2 = \sigma^2 - \sigma^2 = 0.$$



Ένας **μη** αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 2

Αν η αναμενόμενη τιμή μ της ΤΜ Χ ΔΕΝ είναι γνωστή, τότε ο εκτιμητής $S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$

ΔΕΝ είναι αμερόληπτος εκτιμητής για το σ^2.

Διαισθητικά, το γεγονός αυτό οφείλεται στο γεγονός πως το άθροισμα λαμβάνει τη μικρότερη τιμή του όταν η απόσταση των παρατηρήσεων μετριέται από τη μέση τους τιμή.

(Εφαρμογή 2, σελ. 98, σχολικό βιβλίο μαθηματικών γενικής παιδείας Γ Λυκείου.)

2. **Να αποδειχτεί ότι η συνάρτηση**

$$f(\lambda) = \sum_{i=1}^{\nu}(x_i - \lambda)^2 = (x_1 - \lambda)^2 + (x_2 - \lambda)^2 + \ldots + (x_\nu - \lambda)^2$$

γίνεται ελάχιστη, όταν $\lambda = \bar{x}$.

Το ερώτημα που ανακύπτει είναι:

Πόσο μεγαλύτερη πρέπει να γίνει η ποσότητα S^2_n, ώστε να προσεγγίζει με ικανοποιητικό τρόπο την άγνωστη διακύμανση σ^2 του πληθυσμού;

Ένας μη αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 2

Αν η αναμενόμενη τιμή μ της ΤΜ Χ ΔΕΝ είναι γνωστή, τότε ο εκτιμητής $S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$

ΔΕΝ είναι αμερόληπτος εκτιμητής για το σ^2.

$$\mathrm{E}[S_n^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu) - (\overline{X}-\mu)\right)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu)^2 - 2(\overline{X}-\mu)(X_i-\mu) + (\overline{X}-\mu)^2\right)\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{n}(\overline{X}-\mu)^2\sum_{i=1}^{n}1\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + \frac{1}{n}(\overline{X}-\mu)^2\cdot n\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\overline{X}-\mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\cdot n\cdot(\overline{X}-\mu) + (\overline{X}-\mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - 2(\overline{X}-\mu)^2 + (\overline{X}-\mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}-\mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \mathrm{E}\left[(\overline{X}-\mu)^2\right]$$

$$= \sigma^2 - \mathrm{E}\left[(\overline{X}-\mu)^2\right] = \left(1-\frac{1}{n}\right)\sigma^2 < \sigma^2.$$

Πηγή: https://en.wikipedia.org/wiki/Bias_of_an_estimator

Ένας (δειγματικός) αμερόληπτος εκτιμητής για το σ^2

Παράδειγμα 3

Αν $S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$ βρήκαμε ότι $E(S_n^2) = (n-1)\frac{\sigma^2}{n}$, από όπου συνάγεται πως το στατιστικό

$$s_n^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

είναι αμερόληπτος εκτιμητής για το σ^2. Πράγματι:

$$E(s_n^2 - \sigma^2) = \frac{n}{n-1}(n-1)\frac{\sigma^2}{n} - \sigma^2 = \sigma^2 - \sigma^2 = 0.$$

Ο $s_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$ είναι ο βασικός τύπος υπολογισμού της διακύμανσης στη γλώσσα R (συνάρτηση var), στο SPSS και στα λογιστικά φύλλα (συνάρτηση VAR)

Ένας συνεπής εκτιμητής για το σ^2

Παράδειγμα 4

Αν $X \sim N(\mu, \sigma^2)$, και σ^2 άγνωστο τότε η ποσότητα $s_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$ αποδεικνύεται ότι είναι συνεπής εκτιμητής του σ^2.

Απόδειξη: Γνωρίζουμε ότι $E(s_n^2) = \sigma^2$. Αρκεί να δείξω ότι $\text{plim}_{n\to\infty} s_n^2 = \sigma^2$.

Έστω $\varepsilon > 0$. Είναι:

$$P\left(\left|s_n^2 - \sigma^2\right| \geq \varepsilon\right) = P\left(\left|s_n^2 - \mu_{s_n^2}\right| \geq \varepsilon\right) \leq \frac{\text{Var}\left(s_n^2\right)}{\varepsilon^2} = \frac{1}{\varepsilon^2(n-1)^2}\text{Var}\left[\sum_{k=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \frac{\sigma^4}{\varepsilon^2(n-1)^2}\text{Var}\left[\sum_{k=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right] = \frac{\sigma^4}{\varepsilon^2(n-1)^2}\text{Var}(Z_n) = \frac{\sigma^4}{\varepsilon^2(n-1)^2}2(n-1) = \frac{2\sigma^4}{\varepsilon^2(n-1)} \to 0,$$

καθώς $n \to \infty$.

Για την απόδειξη χρησιμοποιήθηκαν:
(α) η ανισότητα του Chebyshev: Αν X Τ.Μ. με $\sigma^2 < \infty$, τότε $P(|X - \mu| > k) \leq \sigma^2/k^2$,
(β) η ιδιότητα $\text{Var}(\lambda X) = \lambda^2\text{Var}(x)$,
(γ) Το γεγονός ότι $Z_n \sim \chi^2(n-1)$ και $\text{Var}(Z_n) = 2(n-1)$.

Κριτήρια

Θεώρημα (Κριτήριο συνέπειας)

Έστω ένας εκτιμητής θ_n, ο οποίος βασίζεται σ' ένα δείγμα μεγέθους n. Αν

$$\lim_{n \to +\infty} E(\theta_n) = \mu$$

και

$$\lim_{n \to +\infty} Var(\theta_n) = 0$$

τότε ο θ_n είναι ένας συνεπής εκτιμητής της θ.

Μία απόδειξη είναι διαθέσιμη εδώ: https://stats.stackexchange.com/questions/17706/how-to-show-that-an-estimator-is-consistent

Θεώρημα (Κριτήριο αμεροληψίας)

Έστω ένας εκτιμητής θ_n, ο οποίος βασίζεται σ' ένα δείγμα μεγέθους n. Αν ο εκτιμητής είναι συνεπής και έχει πεπερασμένη διακύμανση ($Var(\theta_n) < +\infty$) τότε ο θ_n είναι αμερόληπτος.

Μία απόδειξη είναι διαθέσιμη εδώ: https://math.stackexchange.com/questions/239146/consistency-and-asymptotically-unbiasedness

Άσκηση

Σε ένα πληθυσμό υπάρχει μία ιδιότητα στα αντικείμενά του με (άγνωστη) πιθανότητα p. Παρατηρούμε n αντικείμενα και καταγράφουμε το πλήθος X μεταξύ αυτών που έχουν την ιδιότητα. Να δείξετε ότι η ποσότητα $p_n = X/n$ είναι αμερόληπτος και συνεπής εκτιμητής της πιθανότητας p.

Τελικά σχόλια

Ένας εκτιμητής μπορεί να είναι αμερόληπτος αλλά όχι συνεπής

π.χ. Αν επιλέξω n παρατηρήσεις από τον πληθυσμό και ορίσω ως εκτιμητή της αναμενόμενης τιμής τον αριθμητικό μέσο T_2 των 2 μεγαλύτερων τιμών τότε αυτός είναι ένας αμερόληπτος ($E(T_2) = \mu$) αλλά όχι συνεπής εκτιμητής ($\text{plim}_{n\to\infty} T_2 \neq \mu$).

Ένας εκτιμητής μπορεί να είναι συνεπής αλλά όχι αμερόληπτος

π.χ. Αν επιλέξω n παρατηρήσεις και ορίσω $T_n = (x_1 + x_2 + \ldots + x_n)/n + 1/n$, τότε ο T_n είναι μεροληπτικός ($\text{Bias}(T_n) = E(T_n - \mu) = 1/n$), αλλά είναι συνεπής (απόδειξη με χρήση του κριτηρίου συνέπειας και του Κ.Ο.Θ.).

Ο εντοπισμός αμερόληπτων και συνεπών εκτιμητών για γενικές κατανομές είναι συνήθως δύσκολη υπόθεση.

π.χ. *Estimation Of Parameters of a Lognormal distribution.* Retrieved April 17, 2023, from https://www.jstor.org/stable/43834395

Εκτιμητές μέγιστης πιθανοφάνειας

Εκτιμητές μέγιστης πιθανοφάνειας

Στην Στατιστική, η εκτίμηση μέγιστης πιθανότητας (MLE) είναι μια μέθοδος εκτίμησης των παραμέτρων μιας άγνωστης κατανομής πιθανότητας, από ένα σύνολο παρατηρούμενων δεδομένων.

Αυτό επιτυγχάνεται με τη μεγιστοποίηση μιας συνάρτησης (μάζας ή πυκνότητας) πιθανότητας, έτσι ώστε, σύμφωνα με το υποτιθέμενο στατιστικό μοντέλο, τα παρατηρούμενα δεδομένα να είναι πιο πιθανά να εμφανιστούν.

Το σημείο στο χώρο των παραμέτρων που μεγιστοποιεί τη συνάρτηση πιθανότητας ονομάζεται **εκτιμητής μέγιστης πιθανότητας.**

Η διαδικασία είναι εφικτή τόσο για συνεχείς όσο και για διακριτές τμ. Η μέθοδος προτάθηκε από τον Fisher το 1912.

Εκτιμητές μέγιστης πιθανοφάνειας

Έστω ότι τα δεδομένα $x = (x_1, \ldots, x_n)$, προέρχονται από μία κατανομή της οποίας γνωρίζουμε το είδος της κατανομής αλλά όχι τις παραμέτρους αυτής. Έστω επίσης πως οι παρατηρήσεις **είναι ανεξάρτητες.**

Αν f η σ.μ.π. ή σ.π.π και θ είναι μία άγνωστη παράμετρος του πληθυσμού τότε

$$x_i \text{ ανεξάρτητες } \rightarrow f_X(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

Ορίζουμε τη συνάρτηση πιθανοφάνειας L ως εξής :

$$L(\theta) = L(X, \theta) = f_X(x_1, \ldots, x_n; \theta)$$

Αναζητούμε την τιμή της παραμέτρου θ που μεγιστοποιεί τη συνάρτηση L. Η τιμή που βρίσκουμε με την παραπάνω διαδικασία ονομάζεται

Εκτιμητής Μέγιστης Πιθανοφάνειας.

Εκτιμητές μέγιστης πιθανοφάνειας

Οι παρατηρήσεις θεωρούνται ανεξάρτητες, άρα

$$f_X(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

Η συνάρτηση πιθανοφάνειας L γράφεται:

$$L(\theta) = L(X, \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

Για να βρούμε την τιμή της παραμέτρου θ που μεγιστοποιεί τη συνάρτηση L, αρκεί να βρούμε την τιμή που μεγιστοποιεί την

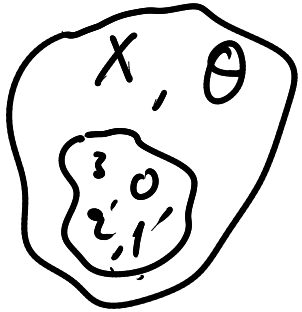
$$l(\theta) = \ln L(\theta) = \Sigma \ln f(x_1; \theta).$$

Ο **εκτιμητής μέγιστης πιθανοφάνειας** προκύπτει από την επίλυση της εξίσωσης

$$\frac{d}{d\theta} l(\theta) = 0 \quad \text{ή} \quad \ln L(X, \theta) = 0.$$

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 1. Έστω X μία ΤΜ με κατανομή



X	0	1	2	3
$f_X(x)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

Για να εκτιμήσουμε την άγνωστη παράμετρο θ, υλοποιούμε δειγματοληψία 10 τιμών και παίρνουμε το δείγμα (3, 0, 2, 1, 3, 2, 1, 0, 2, 1). Να βρεθεί η τιμή της παραμέτρου θ με τη μέθοδο του Εκτιμητή Μέγιστης Πιθανοφάνειας.

$$L(\theta) = f(3,0,2,\dots,1;\theta) = f(3;\theta)\cdot f(0;\theta)\cdot\dots\cdot f(1;\theta) =$$

$$= \frac{1-\theta}{3}\cdot\frac{2\theta}{3}\cdot\dots\cdot\frac{\theta}{3} = \left(\frac{2\theta}{3}\right)^2\cdot\left(\frac{\theta}{3}\right)^3\cdot\left(\frac{2(1-\theta)}{3}\right)^3\cdot\left(\frac{1-\theta}{3}\right)^2 = \frac{2^5}{3^{10}}\cdot\theta^5\cdot(1-\theta)^5.$$

$$\ell(\theta) = \ln L(\theta) = \ln \frac{2^5}{3^{10}} + 5 \ln \theta + 5 \cdot \ln(1-\theta).$$

$$\ell'(\theta) = \frac{5}{\theta} - \frac{5}{1-\theta}, \quad \ell'(\theta) = 0 \Leftrightarrow \theta = \frac{1}{2}$$

θ	0		$\frac{1}{2}$		1
$\ell'(\theta)$		$+$	0	$-$	
$\ell(\theta)$		$\nearrow$	MAX	$\searrow$	

Το $\theta = \frac{1}{2}$ είναι ο Ε.Μ.Π. (MLE).

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 1.

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 2. Έστω ότι η τ.μ. X μετράει το χρόνο μεταξύ αφίξεων σε μία διαδικασία Poisson και πως X ~ Exp(λ). Παρακολουθούμε δείγμα 4 τιμών της X και βρίσκουμε τις τιμές 1.23, 3.32, 1.98, 2.12. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου λ.

$1.23,\ 3.32,\ 1.98,\ 2.12 \qquad X \sim Exp(\lambda),\ f(x) = \lambda e^{-\lambda x},\ x \geq 0.$

$$L(\lambda) = f((1.23, 3.32, 1.98, 2.12); \lambda) = f(1.23;\lambda) \cdot f(3.32;\lambda) \cdot f(1.98;\lambda) f(2.12;\lambda)$$

$$= \lambda \cdot e^{-1.23\cdot\lambda} \cdot \lambda \cdot e^{-3.32\lambda} \cdot \lambda e^{-1.98\lambda} \cdot \lambda e^{-2.12\lambda} = \lambda^4 \cdot e^{-8.65\lambda}$$

$$\ell(\lambda) = \ln L(\lambda) = 4\ln\lambda - 8.65\lambda. \Rightarrow \ell'(\lambda) = \frac{4}{\lambda} - 8.65$$

$$\ell(\lambda) = 0 \Leftrightarrow \lambda = \frac{4}{8.65} = \frac{4}{1.23+3.32+1.98+2.12}.$$

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 2 (γενίκευση). Έστω ότι η τ.μ. X μετράει το χρόνο μεταξύ αφίξεων σε μία διαδικασία Poisson και πως X ~ Exp(λ). Παρακολουθούμε δείγμα n τιμών της X και βρίσκουμε τις τιμές X_1, X_2, …, X_n,. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου λ.

Μέγιστη Πιθανοφάνεια $\not\rightarrow$ Αμεροληψία

Παράδειγμα 1 (με αφορμή την άσκηση 2)
Έστω X μία μεταβλητή που γνωρίζουμε ότι ακολουθεί την Exp(λ) αλλά δεν γνωρίζουμε το λ (= πλήθος αφίξεων / μονάδα χρόνου). Για το λόγο αυτό παρατηρούμε n διαφορετικές διάρκειες μεταξύ αφίξεων $X_1, X_2, \ldots, X_n$ και εκτιμούμε τον ρυθμό λ από τον τύπο:

$$\hat{\lambda} = \frac{n}{X_1 + X_2 + \ldots + X_n}$$

Ο εκτιμητής αυτός **δεν είναι αμερόληπτος**. Πράγματι, για n = 1, είναι

$$E(\hat{\lambda}) = E\left(\frac{1}{X_1}\right) = \int_0^{+\infty} \frac{\lambda}{x} e^{-\lambda/x} dx = +\infty$$

ενώ για n > 1 είναι $X_1 + X_2 + \ldots + X_n \sim$ Γάμμα(n, λ), άρα $1/(X_1 + X_2 + \ldots + X_n) \sim$ Inverse Γάμμα(n, λ) και

$$E(\hat{\lambda}) = E\left(\frac{n}{X_1 + X_2 + \ldots + X_n}\right) = n\frac{\lambda}{n-1} \neq \lambda .$$

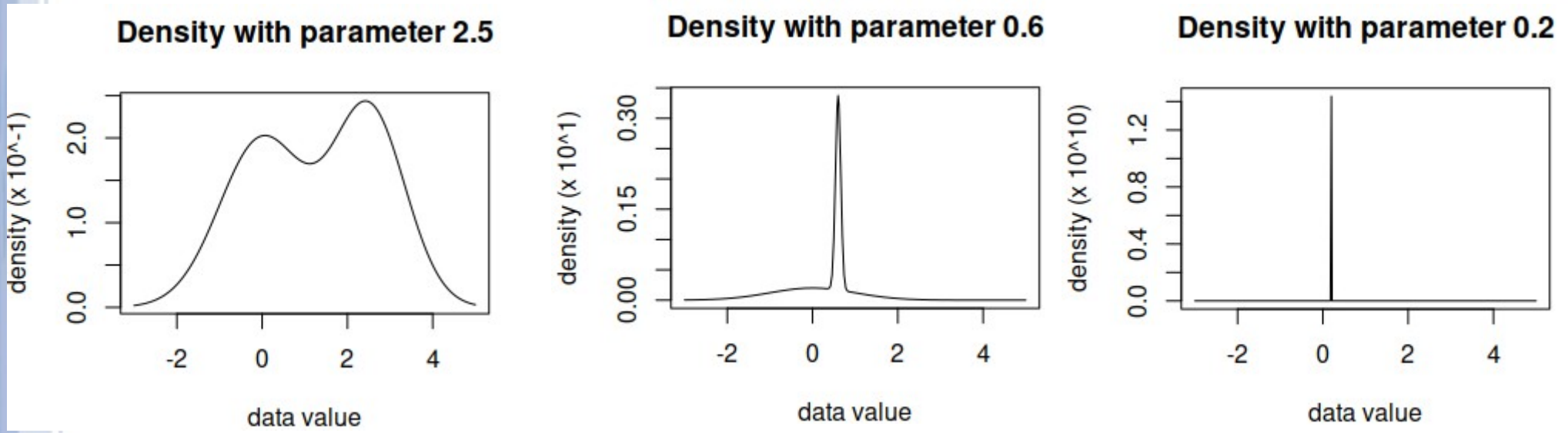
Ιδιαίτερα, συμπεραίνουμε, ότι ο $\hat{\lambda} = \frac{n-1}{X_1 + X_2 + \ldots + X_n}$ είναι ένας αμερόληπτος εκτιμητής του λ.

Σημείωση: Αν X ~ Inverse Γαμμα(α, β) τότε $E(X^k) = \beta^k \Gamma(\alpha - k) / \Gamma(\alpha)$.

Μέγιστη Πιθανοφάνεια $\not\rightarrow$ Συνέπεια

Παράδειγμα 2

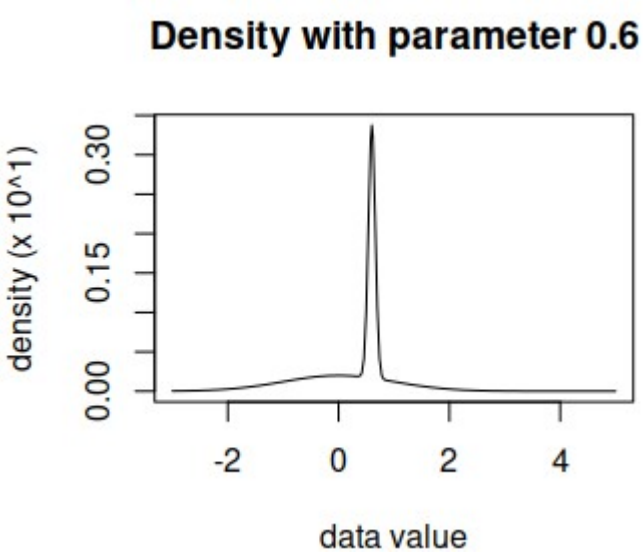
Έστω X τμ για την οποία γνωρίζουμε ότι $X \sim 1/2N(0,1) + 1/2N\left(t, e^{-\frac{2}{t^2}}\right)$.

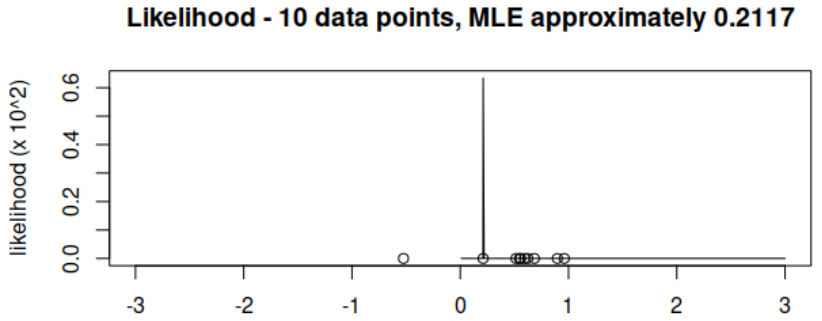


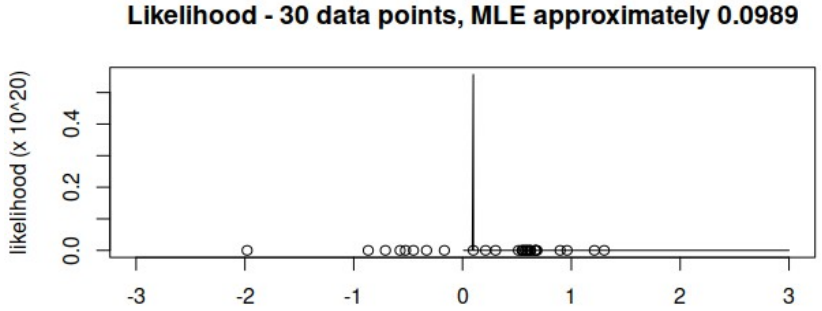
Πηγή: https://radfordneal.wordpress.com/2008/08/09/inconsistent-maximum-likelihood-estimation-an-ordinary-example/

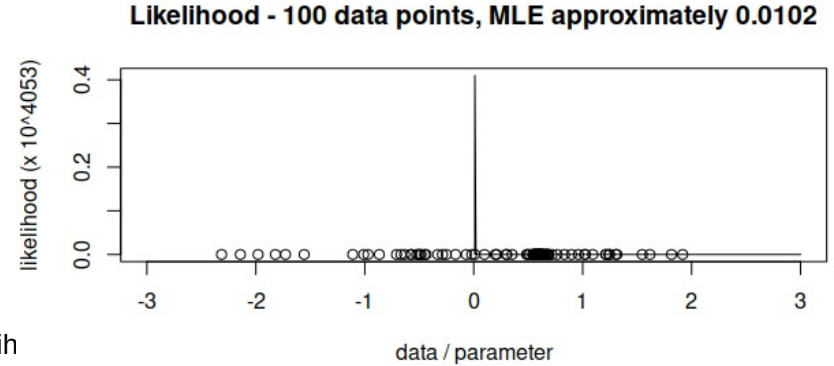
Μέγιστη Πιθανοφάνεια $\not\rightarrow$ Συνέπεια

Παράδειγμα 2
Για t = 0,6, και $\{x_i\}_{i\leq n}$, n = 10 ή 30 ή 100 παρατηρήσεις της X, ο Ε.Μ.Π. της παραπάνω κατανομής **δεν είναι συνεπής**.









Πηγή: https://radfordneal.wordpress.com/2008/08/09/inconsistent-maximum-likelih

Μέγιστη Πιθανοφάνεια $\not\rightarrow$ Μοναδικότητα

Παράδειγμα 3

Αν X ~ Laplace(μ, β) με $f(x) = \frac{1}{2\beta}\exp\left(-\frac{|x-\mu|}{\beta}\right)$, με άγνωστες παραμέτρους μ, β και $x_1, x_2, \ldots, x_n$, ένα δείγμα τιμών, τότε αποδεικνύεται ότι ο εκτιμητής μέγιστης πιθανοφάνειας για την παράμετρο μ είναι η διάμεσος του δείγματος(1).

Καθώς η διάμεση τιμή δεν είναι μοναδική όταν το n είναι άρτιος, καταλαβαίνουμε ότι ο εκτιμητής δεν ορίζεται με μοναδικό τρόπο.

(1) Μία απόδειξη είναι διαθέσιμη εδώ:
https://math.stackexchange.com/questions/240496/finding-the-maximum-likelihood-estimator

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 3. Έστω X μία ΤΜ με κατανομή

$$f(x) = \frac{1}{2\beta} e^{-\frac{|x|}{\beta}}$$

$X_1, \dots, X_n$

Επιλέγουμε δείγμα n τιμών από την κατανομή της X. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου β.

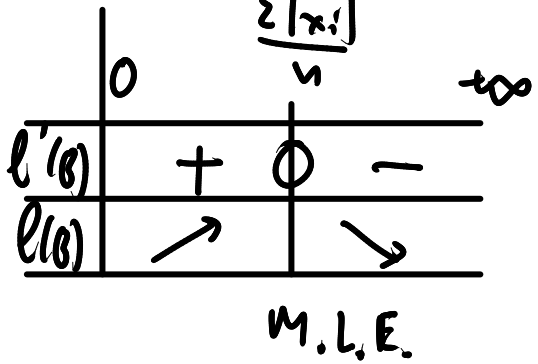
$$L(\beta) = f(x_1, \dots, x_n; \beta) = f(x_1) \cdots f(x_n) = \frac{1}{2\beta} e^{-\frac{|x_1|}{\beta}} \cdots \frac{1}{2\beta} e^{-\frac{|x_n|}{\beta}}$$

$$= \left(\frac{1}{2\beta}\right)^n e^{-\frac{|x_1| + \cdots + |x_n|}{\beta}}$$

$$l(\beta) = \ln L(\beta) = \ln\left(\frac{1}{2\beta}\right)^n - \frac{\sum_{i=1}^{n} |x_i|}{\beta} \Rightarrow l'(\beta) = -\frac{n}{\beta} + \frac{\sum |x_i|}{\beta^2}$$

$-n \ln(2\beta)$

$$\ell'(\beta) = 0 \iff \beta = \frac{\sum |x_i|}{n} .$$



Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 3.

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 4. Μία ΤΜ Χ, ακολουθεί την κατανομή Pareto με σ.π.π.

$$f(x) = \theta x^{-\theta-1},\ x \geq 1,\ \theta > 1.$$

Επιλέγουμε δείγμα n τιμών από την κατανομή της Χ. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου θ.

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 4.

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 5. Για μία ΤΜ X, γνωρίζουμε ότι X ~ U(0, θ). Επιλέγουμε δείγμα n τιμών από την κατανομή της X. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου θ.

$X \sim U(0,\theta)$, $f(x) = \frac{1}{\theta}$, $0 \le x_1, x_2, \dots, x_n \le \theta$.

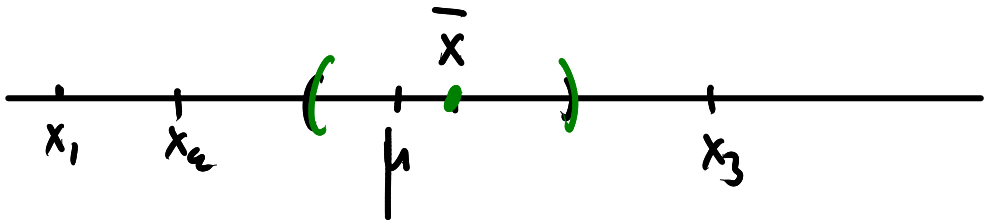
$L(\theta) = f(x_1, \dots, x_n) = f(x_1) \cdots f(x_n) = \frac{1}{\theta} \cdots \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n$.

$\ell(\theta) = \ln L(\theta) = -n \ln \theta$.

$\ell'(\theta) = -\frac{n}{\theta}$.

$MLE = \max\{x_1, x_2, \dots, x_n\}$.

Διάστημα εμπιστοσύνης (confidence interval)



Ποσοστιαία σημεία κανονικής κατανομής

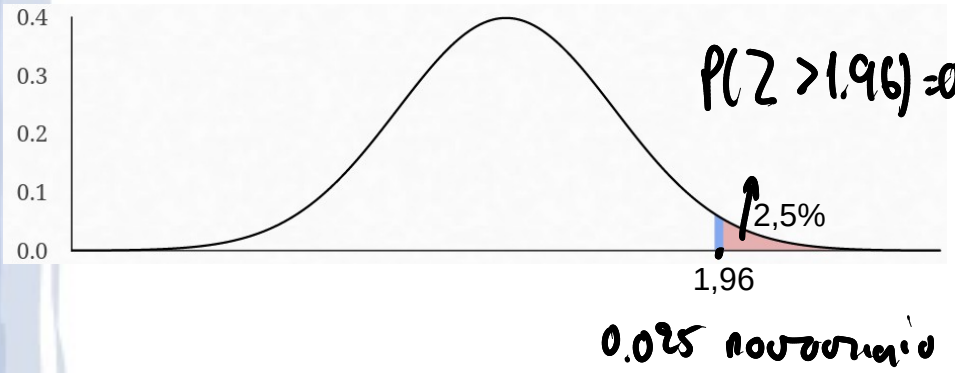
Έστω $Z \sim N(0, 1)$, $0 < \alpha < 1$, και z_α ο αριθμός για τον οποίο

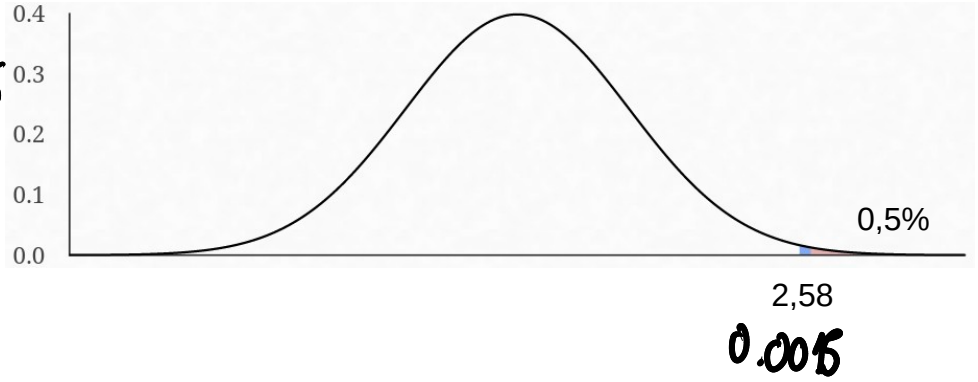
$$\mathbf{P(Z > z_\alpha) = \alpha \text{ ή } \Phi(z_\alpha) = 1 - \alpha.}$$

Το z_α ονομάζεται α – ποσοστιαίο σημείο της κανονικής κατανομής. Ενδεικτικά:

Αν $\alpha = 0{,}025$, τότε $z_{0,025}$ είναι ο αριθμός για τον οποίον $\Phi(z_{0,025}) = 0{,}975$ ή $\mathbf{z_{0,025} = 1{,}96.}$

Αν $\alpha = 0{,}005$, τότε $z_{0,005}$ είναι ο αριθμός για τον οποίον $\Phi(z_{0,005}) = 0{,}995$ ή $\mathbf{z_{0,005} = 2{,}58.}$





Κατανομή αριθμητικού μέσου

Έστω ότι μετρούμε n > 30, τιμές $x_1, x_2, \ldots, x_n$, από έναν πληθυσμό με άγνωστη αναμενόμενη τιμή μ και τυπική απόκλιση σ.

Ο αριθμητικός μέσος των τιμών είναι ο αριθμός $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$ και είναι **μία σημειακή εκτίμηση** για την άγνωστη μέση τιμή του πληθυσμού.

Κάθε μία από τις τιμές που παρατηρήσαμε είναι το αποτέλεσμα ενός ανεξάρτητου πειράματος δειγματοληψίας στον ίδιο πληθυσμό (ανεξάρτητο = καλής ποιότητας δειγματοληψία). Πιο συγκεκριμένα, η τιμή x_i που παρατηρήσαμε είναι μία από τις πιθανές τιμές που παίρνει μία τυχαία μεταβλητή, την οποία ας την ονομάσουμε X_i, i = 1, 2, …, n. Με τον τρόπο αυτό ο αριθμητικός μέσος που υπολογίσαμε αποτελεί μία από τις πιθανές τιμές που μπορεί να πάρει η τυχαία μεταβλητή

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Αναδεικνύεται με φυσικό τρόπο το ερώτημα:

Ποια είναι η κατανομή της συνεχούς μεταβλητής $\bar{X}$;

Κατανομή αριθμητικού μέσου

Καθώς, n > 30 και οι τ.μ. X_i, i = 1, 2, …, n, είναι ισόνομες (αφού λαμβάνονται από τον ίδιο πληθυσμό) και ανεξάρτητες, το Κεντρικό Οριακό Θεώρημα μας εξασφαλίζει ότι: $X_1 + X_2 + \ldots + X_n \sim N(n \cdot \mu, n \cdot \sigma^2)$, ή ισοδύναμα ότι:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Δηλαδή προκύπτει το αξιοσημείωτο συμπέρασμα, πως:

Ανεξάρτητα από την κατανομή που ακολουθεί ο πληθυσμός μέσα από τον οποίο λαμβάνουν τιμές οι X_1, X_2, …, X_n, η τυχαία μεταβλητή του αριθμητικού μέσου, για n αρκετά μεγάλο, ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2 / n)$,

Δραστηριότητα
Επαληθεύστε ότι αν $X \sim N(\mu, \sigma^2)$, και $Z = (X - \mu) / \sigma$, τότε $Z \sim N(0, 1)$.

Αναπαράσταση: https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

Διάστημα Εμπιστοσύνης: σ γνωστό

Το συμπέρασμα που καταλήξαμε: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ δίνει τη δυνατότητα να κάνουμε μία **εκτίμηση σε διάστημα για την (άγνωστη) μέση τιμή μ του πληθυσμού.** Πράγματι, από τη θεωρία που συνοδεύει την κανονική κατανομή, γνωρίζουμε ότι:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1).$$

$X \sim N(\mu, \sigma^2)$

$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$

Αν α > 0, θεωρούμε τον αριθμό $z_{\alpha/2}$, για τον οποίο $P(Z > z_{\alpha/2}) = \alpha/2$. Τότε, θα είναι:

$$P\left(-z_{\alpha/2} < \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} < z_{\alpha/2}\right) = 1 - \alpha \Leftrightarrow P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha.$$

Τι σημαίνει η τελευταία ισότητα; Η απάντηση είναι:
Εάν E(X) = μ και η διαδικασία λήψης δείγματος μεγέθους n, επαναληφθεί πολλές φορές, τότε σε κάθε 100 διαστήματα της μορφής

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$$

στα (1 – α)·100 από αυτά θα βρίσκεται μέσα η πραγματική μέση τιμή μ του πληθυσμού.

Διάστημα Εμπιστοσύνης: σ γνωστό

Το διάστημα

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

ονομάζεται **διάστημα εμπιστοσύνης** (confidence interval) για την άγνωστη μέση τιμή του πληθυσμού.

Οι δύο πιο συνηθισμένες επιλογές για το α είναι α = 0,05 και α = 0,01.

Αν α = 0,05 τότε $z_{0,025} = 1,96$ και βρίσκουμε το

95% διάστημα εμπιστοσύνης $\left(\overline{X} - 1,96 \frac{\sigma}{\sqrt{n}}, \overline{X} + 1,96 \frac{\sigma}{\sqrt{n}} \right)$

Αν α = 0,01 τότε $z_{0,005} = 2,58$ και βρίσκουμε το

99% διάστημα εμπιστοσύνης $\left(\overline{X} - 2,58 \frac{\sigma}{\sqrt{n}}, \overline{X} + 2,58 \frac{\sigma}{\sqrt{n}} \right)$

Αναπαράσταση: https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm

Διάστημα Εμπιστοσύνης: σ γνωστό

Η ποσότητα $SE = \frac{\sigma}{\sqrt{n}}$ είναι η τυπική απόκλιση της κατανομής της δειγματικής μέσης τιμής και ονομάζεται **τυπικό σφάλμα** της δειγματικής κατανομής.

Δηλαδή, μπορούμε να γράψουμε

$$\overline{X} \sim N(\mu, SE^2) \Leftrightarrow \frac{(\overline{X} - \mu)}{SE} \sim N(0, 1).$$

Επιπλέον:

95% διάστημα εμπιστοσύνης $(\overline{X} - 1{,}96\,SE, \overline{X} + 1{,}96\,SE)$

99% διάστημα εμπιστοσύνης $(\overline{X} - 2{,}58\,SE, \overline{X} + 2{,}58\,SE)$

Διάστημα Εμπιστοσύνης: Συνήθη σφάλματα

Λανθασμένες ερμηνείες για το 95% διάστημα εμπιστοσύνης $\left(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right)$

- Επίπεδο εμπιστοσύνης 95% δεν σημαίνει ότι στο ένα διάστημα που υπολογίζουμε από το δείγμα μας, υπάρχει 95% πιθανότητα το διάστημα να καλύπτει την παράμετρο πληθυσμού. Σε αυτό το ένα που υπολογίζουμε είτε θα είναι είτε δεν θα είναι. Αν όμως υπολογίσουμε 100 τότε περιμένουμε τα 95 από αυτά να καλύπτουν την άγνωστη μέση τιμή μ.

- Ένα 95% διάστημα εμπιστοσύνης δεν περιέχει το 95% των τιμών του δείγματος.

Διάστημα Εμπιστοσύνης

Άσκηση 1
Υποθέτουμε ότι οι βαθμολογίες στις εξετάσεις κατανέμονται κανονικά. Δεν γνωρίζουμε τη μέση τιμή αλλά γνωρίζουμε πως η τυπική απόκλιση είναι σ = 3 μονάδες. Σε δείγμα 36 γραπτών βρέθηκε μέση βαθμολογία ίση με 68 μονάδες. Να βρεθεί το 95% και το 99% διάστημα εμπιστοσύνης για την άγνωστη μέση βαθμολογία όλων των φοιτητών.

$\mu = ;\ \sigma = 3,\ n = 36,\ \bar{x} = 68$

95% δ.ε. $\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = \left(68 - 1.96\frac{3}{6}, 68 + 1.96\frac{3}{6}\right)$

$= (67.02, 68.98).$

99% δ.ε. $\left(\bar{x} - 2.59\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.59\frac{\sigma}{\sqrt{n}}\right) = (66.705, 69.295)$

Διάστημα Εμπιστοσύνης: σ άγνωστο

Στην πράξη, συνήθως δεν γνωρίζουμε την τυπική απόκλιση σ του πληθυσμού. Ως εκ τούτου, δεν είναι δυνατόν να υπολογίσουμε το διάστημα εμπιστοσύνης ως

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2},\ \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$$

σ: διακύμανση πληθυσμού

Ωστόσο, στη θέση της άγνωστης τυπικής απόκλισης του πληθυσμού μπορούμε να χρησιμοποιήσουμε **τη δειγματική μέση απόκλιση s**. Στην περίπτωση αυτή, αποδεικνύεται ότι:

$$\frac{\sqrt{n}(\overline{X} - \mu)}{s} \sim t(n-1),$$

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

και ανάλογα προκύπτει ότι το 95% διάστημα εμπιστοσύνης είναι το

$$\left(\overline{X} - \frac{s}{\sqrt{n}} t_{n-1,\ \alpha/2},\ \overline{X} + \frac{s}{\sqrt{n}} t_{n-1,\ \alpha/2}\right)$$

Όπου t(n – 1) είναι η κατανομή Student με n – 1 βαθμούς ελευθερίας και $t_{n-1;\ \alpha/2}$, το αντίστοιχο ποσοστιαίο σημείο.

Ποσοστιαία σημεία t(n) κατανομής

Έστω $X \sim t(n)$, $0 < \alpha < 1$, και $t_{n,\alpha}$ ο αριθμός για τον οποίο

$$\mathbf{P(X > t_{n,\alpha}) = \alpha \text{ ή } P(X < t_{n,\alpha}) = 1 - \alpha.}$$

Το $t_{n,\alpha}$ ονομάζεται **α – ποσοστιαίο σημείο της t(n) κατανομής**. Οι ακριβείς τιμές $\mathbf{t_{n,\alpha}}$ μπορούν να βρεθούν από πίνακες είτε με χρήση υπολογιστή.

Αν n > 30, τότε αρκεί η χρήση της κανονικής κατανομής γιατί αποδεικνύεται ότι:

$$\mathbf{\text{αν } n > 30, \text{ τότε } t_{n,\alpha} \approx z_{\alpha}.}$$

Σημείωση
Μία απόδειξη για τον ισχυρισμό $t_{n,\alpha} \approx z_{\alpha}$ είναι διαθέσιμη εδώ:
https://math.stackexchange.com/questions/3240536/convergence-of-students-t-distribution-to-a-standard-normal

T-Table Distribution

df/alpha	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.706200	31.820520	63.656740	636.619200
2	0.288675	0.816497	1.885618	2.919986	4.302650	6.964560	9.924840	31.599100
3	0.276671	0.764892	1.637744	2.353363	3.182450	4.540700	5.840910	12.924000
4	0.270722	0.740697	1.533206	2.131847	2.776450	3.746950	4.604090	8.610300
5	0.267181	0.726687	1.475884	2.015048	2.570580	3.364930	4.032140	6.868800
6	0.264835	0.717558	1.439756	1.943180	2.446910	3.142670	3.707430	5.958800
7	0.263167	0.711142	1.414924	1.894579	2.364620	2.997950	3.499480	5.407900
8	0.261921	0.706387	1.396815	1.859548	2.306000	2.896460	3.355390	5.041300
9	0.260955	0.702722	1.383029	1.833113	2.262160	2.821440	3.249840	4.780900
10	0.260185	0.699812	1.372184	1.812461	2.228140	2.763770	3.169270	4.586900
11	0.259556	0.697445	1.363430	1.795885	2.200990	2.718080	3.105810	4.437000
12	0.259033	0.695483	1.356217	1.782288	2.178810	2.681000	3.054540	4.317800
13	0.258591	0.693829	1.350171	1.770933	2.160370	2.650310	3.012280	4.220800
14	0.258213	0.692417	1.345030	1.761310	2.144790	2.624490	2.976840	4.140500
15	0.257885	0.691197	1.340606	1.753050	2.131450	2.602480	2.946710	4.072800
16	0.257599	0.690132	1.336757	1.745884	2.119910	2.583490	2.920780	4.015000
17	0.257347	0.689195	1.333379	1.739607	2.109820	2.566930	2.898230	3.965100
18	0.257123	0.688364	1.330391	1.734064	2.100920	2.552380	2.878440	3.921600
19	0.256923	0.687621	1.327728	1.729133	2.093020	2.539480	2.860930	3.883400
20	0.256743	0.686954	1.325341	1.724718	2.085960	2.527980	2.845340	3.849500
21	0.256580	0.686352	1.323188	1.720743	2.079610	2.517650	2.831360	3.819300
22	0.256432	0.685805	1.321237	1.717144	2.073870	2.508320	2.818760	3.792100
23	0.256297	0.685306	1.319460	1.713872	2.068660	2.499870	2.807340	3.767600
24	0.256173	0.684850	1.317836	1.710882	2.063900	2.492160	2.796940	3.745400
25	0.256060	0.684430	1.316345	1.708141	2.059540	2.485110	2.787440	3.725100
26	0.255955	0.684043	1.314972	1.705618	2.055530	2.478630	2.778710	3.706600
27	0.255858	0.683685	1.313703	1.703288	2.051830	2.472660	2.770680	3.689600
28	0.255768	0.683353	1.312527	1.701131	2.048410	2.467140	2.763260	3.673900
29	0.255684	0.683044	1.311434	1.699127	2.045230	2.462020	2.756390	3.659400
30	0.255605	0.682756	1.310415	1.697261	2.042270	2.457260	2.750000	3.646000

Διάστημα Εμπιστοσύνης

Άσκηση 2

Τον Οκτώβριο του 2008, εξετάστηκαν 20 νεογέννητα βρέφη στις Ηνωμένες Πολιτείες ως προς το πλήθος βιομηχανικών ενώσεων, ρύπων και άλλων χημικών ουσίες που συνδέονται με την τοξικότητα του εγκεφάλου και του νευρικού συστήματος, την τοξικότητα του ανοσοποιητικού συστήματος, την αναπαραγωγική τοξικότητα και προβλήματα γονιμότητας. Συνολικά εξετάστηκαν 430 ουσίες. Στο αίμα ομφάλιου λώρου των 20 βρεφών βρέθηκαν οι εξής ουσίες (σε πλήθος):

79 145 147 160 116 100 159 151 156 126 137 83 156 94 121 144 123 114 139 99

Να βρεθεί 95% διάστημα εμπιστοσύνης για το μέσο πλήθος ουσιών του συνόλου των νεογέννητων βρεφών.

Δίνεται ότι η μέση τιμή του δείγματος είναι 127,45, η τυπ. απόκλιση s = 25,965 και ότι $t_{19;0.025} = 2,093$

Διάστημα Εμπιστοσύνης

Άσκηση 3

Δεκαπέντε εθελοντές συμμετείχαν σε μία μελέτη για την ανακούφιση του πόνου που οφείλεται στο βελονισμό, Τα αποτελέσματα ήταν τα εξής:

8,6; 9,4; 7,9; 6,8; 8,3; 7,3; 9,2; 9,6; 8,7; 11,4; 10,3; 5,4; 8,1; 5,5; 6,9

Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για τη μέση ανακούφιση που θα δηλωθεί στο σύνολο του πληθυσμού στον οποίο θα εφαρμοστεί αυτή η μέθοδος.

Δίνεται ότι η μέση τιμή του δείγματος είναι 8,23, η τυπ, απόκλιση s = 1,67 και ότι $t_{14;0,025} = 2,140$

Διάστημα Εμπιστοσύνης για Ποσοστό

Διάστημα Εμπιστοσύνης για Ποσοστό

Έστω ότι ένα πείραμα διεξάγεται με πιθανότητα επιτυχίας p την οποία δεν γνωρίζουμε. Εκτελούμε n φορές το πείραμα και βρίσκουμε X επιτυχίες. Μία σημειακή εκτίμηση του άγνωστου ποσοστού είναι

$$\hat{p} = \frac{X}{n}$$

Γνωρίζουμε όμως, ότι αν X = {πλήθος από επιτυχίες στις n επαναλήψεις ενός πειράματος} τότε $X \sim B(n,p)$. Επιπλέον, αν $n > 30$, $np > 5$, $nq > 5$, τότε $X \sim N(np, npq)$ ή $X/n \sim N(p, pq/n)$ ή

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right)$$

Το τελευταίο μας δίνει τη δυνατότητα να υπολογίσουμε με ανάλογο τρόπο, το α – διάστημα εμπιστοσύνης για το άγνωστο ποσοστό.

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right) \Leftrightarrow \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{pq}} \sim N(0, 1).$$

Αν $\alpha > 0$, θεωρούμε τον αριθμό $z_{\alpha/2}$, για τον οποίο $P(Z > z_{\alpha/2}) = \alpha/2$. Τότε, θα είναι:

$$P\left(-z_{\alpha/2} < \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{pq}} < z_{\alpha/2}\right) = 1 - \alpha \Leftrightarrow P\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2} < p < \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2}\right) = 1 - \alpha.$$

Διάστημα Εμπιστοσύνης για Ποσοστό

Το διάστημα $\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2},\ \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2}\right)$ ονομάζεται (1 – α) – διάστημα εμπιστοσύνης για το άγνωστο ποσοστό του πληθυσμού. Δηλαδή:

Εάν η διαδικασία λήψης δείγματος μεγέθους n, επαναληφθεί πολλές φορές, τότε σε κάθε 100 διαστήματα της μορφής

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

στα (1 – α)·100 από αυτά θα βρίσκεται μέσα το πραγματικό ποσοστό p του πληθυσμού.

Οι δύο περιπτώσεις που εμφανίζονται συνήθως είναι αυτές του 95% και του 99% δ.ε.

$$95\%\ \delta.\varepsilon.: \left(\hat{p} - 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$99\%\ \delta.\varepsilon.: \left(\hat{p} - 2{,}58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 2{,}58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Διάστημα Εμπιστοσύνης για Ποσοστό

Άσκηση 3
Μία εταιρεία έρευνας αγοράς θέλει να υπολογίσει το ποσοστό των ενηλίκων που ζουν σε μια μεγάλη πόλη και έχουν κινητά τηλέφωνα. Πεντακόσιοι τυχαία επιλεγμένοι ενήλικες κάτοικοι αυτής της πόλης ερευνώνται για να διαπιστωθεί εάν έχουν κινητά τηλέφωνα. Από τα 500 άτομα που συμμετείχαν στην έρευνα, τα 421 απάντησαν ναι - είναι κάτοχοι κινητών τηλεφώνων. Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για την πραγματική αναλογία των ενηλίκων κατοίκων αυτής της πόλης που έχουν κινητά τηλέφωνα.

Διάστημα Εμπιστοσύνης για Ποσοστό

Άσκηση 4
Σε δείγμα 250 τυχαία επιλεγμένων ατόμων, οι 98 ανέφεραν ότι κατέχουν tablet. Χρησιμοποιώντας ένα επίπεδο εμπιστοσύνης 95%, υπολογίστε ένα διάστημα εμπιστοσύνης για το πραγματικό ποσοστό των ατόμων που κατέχουν tablet.

Μέγεθος Δείγματος

Μέγεθος Δείγματος για εκτίμηση μέσης τιμής

Παρατηρούμε ότι το πιθανοθεωρητικό σφάλμα του διαστήματος εμπιστοσύνης

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

είναι ίσο με $E = \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$. Από τον τύπο αυτό, αν θεωρήσουμε δεδομένο το σφάλμα E,

μπορούμε να λύσουμε ως προς το μέγεθος του δείγματος και να βρούμε

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

Ο τύπος αυτός μπορεί να χρησιμοποιηθεί για τον εντοπισμό του μεγέθους δείγματος που χρειάζεται ώστε το μέγιστο σφάλμα μεταξύ του αριθμητικού μέσου του δείγματος και της (άγνωστης) μέσης τιμής να μην ξεπερνάει το E με (1 – α)% πιθανότητα (όπως αυτή εκφράζεται με το (1 – α) δ.ε.).

Μέγεθος Δείγματος για εκτίμηση αναλογίας

Αντίστοιχα, αν γίνεται προσπάθεια εκτίμησης αναλογίας στον πληθυσμό, υπολογίζουμε

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

Καθώς, για x ϵ [0, 1], είναι x(1 – x) ≤ ¼, συμπεραίνουμε ότι $n \leq \left(\frac{z_{\alpha/2}}{2E}\right)^2$.

Συνεπώς, αν το μέγιστο σφάλμα εκτίμησης είναι E:
(α) Αν υπάρχει μία αξιόπιστη εκτίμηση σχετικά την άγνωστη τιμή της αναλογίας, τότε μπορούμε να υπολογίσουμε το μέγεθος δείγματος από τον τύπο

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

(β) Αν δεν υπάρχει αξιόπιστη εκτίμηση σχετικά με την άγνωστη αναλογία, τότε η επιλέγεται η πιο συντηρητική επιλογή:

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

Μέγεθος Δείγματος

Άσκηση 5
Θέλουμε να εκτιμήσουμε το ποσοστό καπνιστών στους κατοίκους της πόλης της Ξάνθης. Πόσο μεγάλο πρέπει να είναι το δείγμα ώστε το σφάλμα να μην ξεπερνάει το 4% με πιθανότητα 95%;

Μέγεθος Δείγματος

Άσκηση 6
Ένας οικονομολόγος επιθυμεί να εκτιμήσει, με εμπιστοσύνη 95%, το ετήσιο εισόδημα των ιδιωτικών υπαλλήλων του Νομού Ξάνθης με μέγιστο σφάλμα 200 ευρώ. Ποιο πρέπει να είναι το μέγεθος του δείγματος που πρέπει να μετρήσει, αν γνωρίζει ότι ο ελάχιστος μισθός είναι 500 ευρώ και ο μέγιστος 2500; (Υποθέτουμε ότι ο μισθός των υπαλλήλων ακολουθεί κανονική κατανομή).