

Πιθανότητες & Στατιστική

Εκτιμητική

Δημήτρης Μητσούδης

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

1 / 26

Τυχαίο δείγμα

Ο βασικός στόχος της Στατιστικής είναι να εξάγουμε συμπεράσματα για ένα συγκεκριμένο σύνολο στοιχείων, το οποίο θα ονομάζουμε **πληθυσμό**, εξετάζοντας ένα σχετικά μικρό υποσύνολο των μελών του, που το λέμε **δείγμα**.

Παράδειγμα: Ένα εργοστάσιο κατασκευής λαμπτήρων πραγματοποιεί στατιστική μελέτη για το χρόνο ζωής των λαμπτήρων.

- 1 **Πληθυσμός:** το σύνολο των λαμπτήρων που κατασκευάζει το εργοστάσιο.
- 2 **Υπό εξέταση χαρακτηριστικό:** ο χρόνος ζωής του λαμπτήρα.

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

2 / 26

Υποθέσεις

Θα θεωρούμε ότι:

- η πιθανοθεωρητική συμπεριφορά του πληθυσμού χαρακτηρίζεται από μια συνάρτηση κατανομής F ,
- η αντίστοιχη ποσοτική μέτρηση του χαρακτηριστικού περιγράφεται από την αντίστοιχη τ.μ. X , με $X \sim F$.

Άρα:

Αν γνωρίζαμε ακριβώς τη συνάρτηση κατανομής F , τότε θα γνωρίζαμε πλήρως και τη συμπεριφορά του πληθυσμού.

Παράδειγμα: Αν η τ.μ. X παριστάνει το χρόνο ζωής ενός λαμπτήρα με αντίστοιχη συνάρτηση κατανομής F , τότε οι τιμές

$$F(x) = P(X \leq x)$$

θα μας έδιναν το **ποσοστό των λαμπτήρων με χρόνο ζωής το πολύ x** .

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

3 / 26



Στην πράξη η F είναι **άγνωστη** και μπορούμε μόνο να παρατηρήσουμε ορισμένες τιμές της τ.μ. X .
(Στο παράδειγμά μας μπορούμε να ελέγξουμε n λαμπτήρες και να παρατηρήσουμε το χρόνο ζωής τους, έστω $X_1, X_2, \dots, X_n$.)

Ορισμός (Τυχαίο Δείγμα)

Αν ένας πληθυσμός έχει συνάρτηση κατανομής F , τότε **τυχαίο δείγμα** είναι ένα σύνολο ανεξάρτητων, ισόνομων τ.μ.

$$X_1, X_2, \dots, X_n$$

με κοινή συνάρτηση κατανομής F . Το n λέγεται **μέγεθος του δείγματος**.

Δ.Α. Μητσούδης

Πιθανότητες & Στατιστική - Τμ. Ναυπηγών

4 / 26

- 1 Στην πράξη, μετά την επιλογή ενός **αντιπροσωπευτικού** δείγματος, οι τ.μ. $X_1, X_2, \dots, X_n$ λαμβάνουν κάποιες συγκεκριμένες (πραγματικές) τιμές

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n.$$

Οι τιμές $x_1, x_2, \dots, x_n$ λέγονται επίσης **τυχαίο δείγμα**, επειδή παριστάνουν τις παρατηρηθείσες τιμές (μετρήσεις) των $X_1, X_2, \dots, X_n$ μετά τη δειγματοληψία.

Στη θεωρία δεν κάνουμε διάκριση μεταξύ $X_1, X_2, \dots, X_n$ και $x_1, x_2, \dots, x_n$, αλλά **τα στατιστικά συμπεράσματα εφαρμόζονται στις μετρήσεις** $x_1, x_2, \dots, x_n$.

- 2 Η Στατιστική χωρίζεται στην (α) **παραμετρική στατιστική**, και στη (β) **μη παραμετρική στατιστική**.

- (α) Στην **μη παραμετρική στατιστική** θεωρούμε ότι η συνάρτηση κατανομής F είναι εντελώς άγνωστη ή ανήκει σε κάποια πολύ γενική κλάση κατανομών (π.χ. η F είναι συνεχής συνάρτηση κατανομής).
- (β) Στην **παραμετρική στατιστική** θεωρούμε ότι μια παράμετρος της κατανομής είναι άγνωστη.

Π.χ., αν η F είναι συνάρτηση κατανομής του χρόνου ζωής των λαμπτήρων, τότε μπορούμε να υποθέσουμε εκ των προτέρων ότι η F είναι εκθετική με **άγνωστη** παράμετρο $\theta > 0$. Τότε

$$F(x) = 1 - e^{-\theta x}, \quad x > 0,$$

επομένως η προσπάθεια μας εστιάζεται **στην εκτίμηση της άγνωστης παραμέτρου θ βάσει της πληροφορίας που θα αντλήσουμε από ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$** .

Εμείς δεν θα ασχοληθούμε με προβλήματα μη παραμετρικής στατιστικής.

- Θα ονομάζουμε **εκτιμήτρια** ($\hat{\theta}$) της άγνωστης παραμέτρου (θ) μια συνάρτηση που εξαρτάται μόνο από τις τιμές του δείγματος που έχουμε επιλέξει και την οποία χρησιμοποιούμε για την εκτίμηση της άγνωστης παραμέτρου θ .
- Θα λέμε **εκτίμηση της παραμέτρου** την αριθμητική τιμή της εκτιμήτριας για ένα συγκεκριμένο δείγμα.

Φυσικά, για μια συγκεκριμένη παράμετρο μπορεί κανείς να ορίσει περισσότερες από μια εκτιμήτριες. Επομένως, οι εκτιμήτριες θα πρέπει να αξιολογούνται με βάση κάποια κριτήρια.

Κριτήρια επιλογής «καλών» εκτιμητριών

- **Αμεροληψία:** Μια εκτιμήτρια $\hat{\theta}$ της άγνωστης παραμέτρου θ θα λέγεται **αμερόληπτη (unbiased)** αν

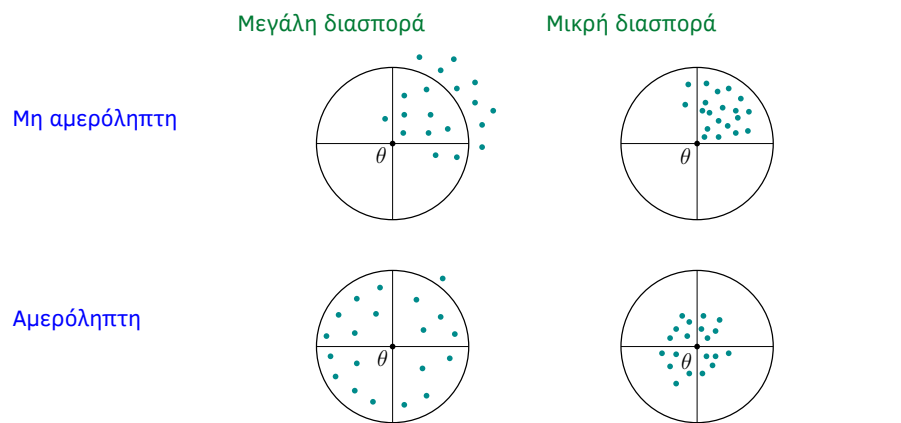
$$E(\hat{\theta}) = \theta.$$

- **Αποτελεσματικότητα:** Η αποτελεσματικότητα αναφέρεται στη διασπορά της εκτιμήτριας και δίνεται συγκριτικά. Μια εκτιμήτρια $\hat{\theta}_1$ της θ είναι πιο **αποτελεσματική (efficient)** από μια άλλη εκτιμήτρια $\hat{\theta}_2$ αν έχει μικρότερη διασπορά, δηλαδή

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2).$$

Εκτός από τα παραπάνω, υπάρχουν και άλλα κριτήρια επιλογής καλών εκτιμητριών (π.χ. συνέπεια), στα οποία εμείς δεν θα αναφερθούμε.

Διαισθητικά, οι έννοιες της αμεροληψίας και της αποτελεσματικότητας απεικονίζονται στο ακόλουθο σχήμα, όπου φανταζόμαστε ότι η εκτιμήτρια $\hat{\theta}$ «στοχεύει» προς το στόχο θ .



Εκτιμήτριες μέσης τιμής και διασποράς

Εκτίμηση μέσης τιμής: Είναι φυσικό να ορίσουμε ως εκτιμήτρια της $\mu = E(X)$ το δειγματικό μέσο

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Εκτίμηση διασποράς: Όπως για τη μέση τιμή έτσι και για τη διασπορά $\sigma^2 = \text{Var}(X)$ η εκτιμήτρια είναι η δειγματική διασπορά

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Αποδεικνύεται ότι οι $\bar{X}$, S^2 είναι «καλές» εκτιμήτριες των μ , σ^2 , αντίστοιχα, (δηλ. είναι αμερόληπτες, αποτελεσματικές, κλπ).

Διαστήματα εμπιστοσύνης

Γενικά, δεν μπορεί κανείς να αναμένει ότι η τιμή της εκτιμήτριας θα συμπίπτει με την άγνωστη παράμετρο. Προφανώς, το κάθε δείγμα θα δίνει τη δική του **σημειακή εκτίμηση**.

Για το λόγο αυτό είναι συχνά χρήσιμο να κατασκευάζουμε ένα **διάστημα εμπιστοσύνης** για την άγνωστη παράμετρο, δηλαδή ένα διάστημα της μορφής $[L, U]$, το οποίο να περιέχει την άγνωστη παράμετρο με «αρκετά μεγάλη πιθανότητα».

Διαστήματα εμπιστοσύνης

Επομένως, ενδιαφερόμαστε να κατασκευάσουμε διαστήματα της μορφής $[L, U]$, έτσι ώστε η **πιθανότητα η άγνωστη παράμετρος θ να βρίσκεται εντός του διαστήματος να είναι ίση με $1 - \alpha$** , όπου $\alpha \in (0, 1)$ είναι μια συνήθως «μικρή» παράμετρος.

Τότε το διάστημα $[L, U]$ λέγεται **διάστημα εμπιστοσύνης (δ.ε.)** για το θ , και η πιθανότητα $1 - \alpha$ λέγεται **συντελεστής εμπιστοσύνης (σ.ε.)** του διαστήματος.

- Για συντομία, λέμε ότι το $[L, U]$ είναι ένα $100(1 - \alpha)\%$ δ.ε. για το θ , και εννοούμε ότι ο σ.ε. του $[L, U]$ είναι $1 - \alpha$.
- Συνήθεις επιλογές για το α είναι 5%, 10%, και 1%, έτσι ώστε ο συντελεστής εμπιστοσύνης $1 - \alpha$ να είναι 95%, 90%, και 99%, αντίστοιχα.

Δ.Ε. για το μέσο μ ενός πληθυσμού με γνωστή διασπορά

Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες παρατηρήσεις από την κανονική κατανομή με μέσο μ και διασπορά σ^2 γνωστή.

Τότε ένα $(1 - \alpha)100\%$ διάστημα εμπιστοσύνης για το μέσο μ είναι το

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]. \quad (1)$$

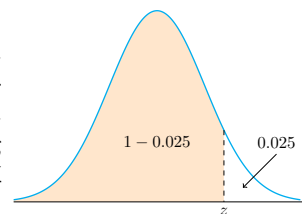
Η ποσότητα $z_{\alpha/2}$ συμβολίζει το άνω εκατοστιαίο σημείο στον οριζόντιο άξονα της τυποποιημένης κανονικής κατανομής στα δεξιά του οποίου υπάρχει εμβαδόν ίσο με $\alpha/2$.

Πώς να υπολογίζετε το $z_{\alpha/2}$

Έστω ότι θέλουμε να κατασκευάσουμε ένα δ.ε. με σ.ε. 95%. Θα πρέπει να είμαστε σε θέση να υπολογίζουμε από τον πίνακα της κανονικής κατανομής την τιμή που αντιστοιχεί στο $z_{\alpha/2}$.

Βήμα 1: $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$.

Βήμα 2: Κάνουμε ένα πρόχειρο γράφημα της καμπύλης. Αναζητούμε στον οριζόντιο άξονα της καμπύλης ένα σημείο z , το οποίο αφήνει δεξιά του εμβαδόν ίσο με 0.025. Άρα αναζητούμε στον πίνακα της κανονικής κατανομής εμβαδόν ίσο με $1 - 0.025 = 0.975$. Το z που αντιστοιχεί σε αυτό το εμβαδόν είναι ίσο με 1.96.



Ανάλογα μπορούμε να υπολογίσουμε το οποιοδήποτε z μας ζητηθεί.

Όπως ήδη είπαμε, συνήθως κατασκευάζουμε διαστήματα με συντελεστή εμπιστοσύνης 90%, 95% και 99%.

Εφαρμόστε την προηγούμενη διαδικασία για να υπολογίσετε τα z που αντιστοιχούν σε δ.ε. με σ.ε. 90% και 99%.

Παράδειγμα 1

Έστω ότι ένα σήμα με τιμή μ μεταδίδεται από τη θέση A και η τιμή που λαμβάνεται στη θέση B ακολουθεί κανονική κατανομή με μέση τιμή μ και διασπορά 4. Δηλαδή, όταν στέλνουμε την τιμή μ από τη θέση A , στη θέση B λαμβάνουμε την τιμή $\mu + N$, όπου ο «θόρυβος» N ακολουθεί κανονική κατανομή με μέσο 0 και διασπορά 4. Για να μειώσουμε το σφάλμα στέλνουμε την ίδια τιμή 9 φορές. Αν οι διαδοχικές τιμές που λαμβάνουμε στη θέση B είναι

5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5,

κατασκευάστε ένα 95% διάστημα εμπιστοσύνης για το μ .

Λύση

Καταρχήν, υπολογίζουμε το δειγματικό μέσο:

$$\bar{X} = \frac{5 + 8.5 + \dots + 10.5}{9} = \frac{81}{9} = 9.$$

Δεδομένου ότι η κατανομή των σημάτων που λαμβάνουμε στη θέση B είναι **κανονική**, και η **διασπορά είναι γνωστή**, το ζητούμενο δ.ε. δίνεται από τον τύπο (1) με σ.ε. $1 - \alpha = 0.95$.

Επομένως, $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$, και από τον πίνακα της τυποποιημένης κανονικής κατανομής βρίσκουμε ότι $z_{\alpha/2} = 1.96$.

Άρα

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = \left[9 - 1.96 \frac{2}{\sqrt{9}}, 9 + 1.96 \frac{2}{\sqrt{9}} \right] = [7.69, 10.31].$$

Επομένως, είμαστε «κατά 95% βέβαιοι» ότι η πραγματική τιμή του μηνύματος είναι μεταξύ των δύο αυτών ορίων.

Δ.Ε. για το μέσο μ ενός πληθυσμού με άγνωστη διασπορά (1)

Έστω $X_1, X_2, \dots, X_n$ **ανεξάρτητες** παρατηρήσεις από πληθυσμό που ακολουθεί $\mathcal{N}(\mu, \sigma^2)$ με διασπορά **σ^2 άγνωστη**.

Τότε ένα $(1 - \alpha)100\%$ διάστημα εμπιστοσύνης για το μέσο μ είναι το

$$\left[\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \right]. \quad (2)$$

S είναι η δειγματική τυπική απόκλιση και $t_{\alpha/2, n-1}$ είναι το άνω $\alpha/2$ εκατοστιαίο σημείο της κατανομής Student με $n - 1$ βαθμούς ελευθερίας (β.ε.).

Λίγα λόγια για την κατανομή Student

Έστω ότι $X \sim \mathcal{N}(\mu, \sigma^2)$. Για ένα δείγμα $X_1, X_2, \dots, X_n$ με δειγματικό μέσο $\bar{X}$ και δειγματική τυπική απόκλιση S , η τ.μ.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (3)$$

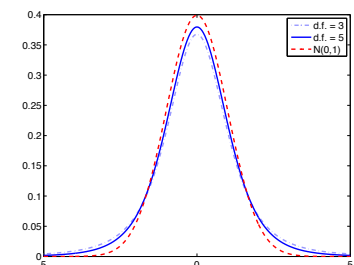
ακολουθεί την κατανομή t του Student με $n - 1$ βαθμούς ελευθερίας.

Για να πάρουμε μια ιδέα για αυτή την κατανομή θα μπορούσαμε να επιλέξουμε πολλά δείγματα μεγέθους n και, στη συνέχεια, να υπολογίσουμε τις αντίστοιχες τιμές του T μέσω της (3), ώστε να σχεδιάσουμε ένα ιστόγραμμα.

Από την άλλη μεριά, υπάρχουν θεωρήματα τα οποία μας παρέχουν τον τύπο της κατανομής αυτής.

Ιδιότητες της κατανομής Student

- Το γράφημά της είναι συμμετρικό ως προς τον y -άξονα, και μοιάζει με «καμπάνα» αλλά με «παχύτερες» ουρές από αυτές της κανονικής κατανομής.
- Καθώς αυξάνουν οι β.ε. η t -κατανομή προσεγγίζει την τυποποιημένη κανονική.



Παράδειγμα 2

Ας θεωρήσουμε και πάλι το Παράδειγμα 1, μόνο που τώρα θα υποθέσουμε ότι η τιμή που λαμβάνεται στη θέση B ακολουθεί κανονική κατανομή με μέση τιμή μ και διασπορά σ^2 (άγνωστη).

Αν οι διαδοχικές τιμές που λαμβάνουμε στη θέση B είναι και πάλι

5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5,

κατασκευάστε ένα 95% διάστημα εμπιστοσύνης για το μ .

n	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.474	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

Αρα

$$\left[\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \right] \\ = \left[9 - 2.306 \frac{3.082}{\sqrt{9}}, 9 + 2.306 \frac{3.082}{\sqrt{9}} \right] = [6.63, 11.37].$$

Επομένως, είμαστε «κατά 95% βέβαιοι» ότι η πραγματική τιμή του μηνύματος είναι μεταξύ των δύο αυτών ορίων.

Λύση

Υπολογίζουμε τη δειγματική μέση τιμή και τη δειγματική διασπορά:

$$\bar{X} = 9 \quad \text{και} \quad S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n \cdot \bar{X}^2 \right) = 9.5.$$

Αρα η δειγματική τυπική απόκλιση είναι $S = 3.082$.

Δεδομένου ότι η κατανομή των σημάτων που λαμβάνουμε στη θέση B είναι κανονική, η διασπορά είναι άγνωστη, και το μέγεθος του δείγματος είναι μικρό, το ζητούμενο δ.ε. δίνεται από τον τύπο (2) με σ.ε. $1-\alpha = 0.95$.

Επομένως, $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$, και από τον πίνακα της κατανομής Student βρίσκουμε ότι $t_{0.025, 8} = 2.306$.

Δ.Ε. για το μέσο μ ενός πληθυσμού με άγνωστη διασπορά (2)

Αν το μέγεθος του δείγματος (n) είναι μεγάλο (στην πράξη συνήθως $n \geq 30$), τότε το κεντρικό οριακό θεώρημα υποδεικνύει ότι ο δειγματικός μέσος ακολουθεί κατά προσέγγιση κανονική κατανομή.

Επομένως, αν το n είναι μεγάλο δεν είναι απαραίτητη η υπόθεση ότι το δείγμα προέρχεται από κανονικό πληθυσμό.

Αρα, αν $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες παρατηρήσεις από κάποια κατανομή με μέσο μ , διασπορά σ^2 άγνωστη, και το n είναι μεγάλο, τότε ένα $(1-\alpha)100\%$ διάστημα εμπιστοσύνης για το μέσο μ είναι το

$$\left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]. \quad (4)$$

Ο τύπος αυτός είναι ίδιος με τον (1) μόνο που στη θέση της πληθυσμιακής τυπ. απόκλισης σ έχει τη δειγματική τυπική απόκλιση S .

Παράδειγμα

Ένα τυχαίο δείγμα από 100 καθηγητές μέσης εκπαίδευσης έδωσε για τις ετήσιες καθαρές αποδοχές δειγματικό μέσο 13400 € και δειγματική τυπική απόκλιση 4120 €. Κατασκευάστε ένα 95% διάστημα εμπιστοσύνης για το μέσο ετήσιο εισόδημα των καθηγητών μ.ε.

Λύση

Δεδομένου ότι **το μέγεθος του δείγματος είναι $n = 100$ (μεγάλο)**, μπορούμε να υποθέσουμε ότι ο δειγματικός μέσος ακολουθεί κατά προσέγγιση κανονική κατανομή και, επομένως, το ζητούμενο δ.ε. δίνεται από τον τύπο (4) με σ.ε. $1 - \alpha = 0.95$.

Εδώ $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$, και από τον πίνακα της τυποποιημένης κανονικής κατανομής βρίσκουμε ότι $z_{\alpha/2} = 1.96$.

Άρα

$$\left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right] = \left[13400 - 1.96 \frac{4120}{\sqrt{100}}, 13400 + 1.96 \frac{4120}{\sqrt{100}} \right] \\ = [12592.48, 14207.52].$$

Επομένως, είμαστε «κατά 95% βέβαιοι» ότι το μέσο ετήσιο εισόδημα είναι μεταξύ των δύο αυτών ορίων.