# ΠΕΡΙΕΧΟΜΕΝΑ

# 11

# ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

## ΒΑΣΙΚΟΙ ΟΡΟΙ

**διάστημα εμπιστοσύνης** (confidence interval): ένα διάστημα που βασίζεται σε παρατηρήσεις ενός δείγματος και είναι καθορισμένο με τέτοιο τρόπο ώστε να υπάρχει μια συγκεκριμένη πιθανότητα ότι θα περιέχει την άγνωστη πραγματική τιμή μιας παραμέτρου (για παράδειγμα, συνηθίζεται ο καθορισμός διαστημάτων εμπιστοσύνης που έχουν 95% πιθανότητα να περιέχουν την πραγματική τιμή)

**επίπεδο εμπιστοσύνης** (confidence level): ο βαθμός εμπιστοσύνης που σχετίζεται με ένα διάστημα εμπιστοσύνης· η πιθανότητα ότι το διάστημα περιέχει την πραγματική τιμή της παραμέτρου

Στο Κεφάλαιο 10 εξηγήσαμε τον τρόπο με τον οποίο μπορούμε να χρησιμοποιούμε τις παρατηρήσεις μιας τυχαίας μεταβλητής για να παίρνουμε πληροφορίες για μια άγνωστη παράμετρο της κατανομής που παράγει τη μεταβλητή. Πάντως, εξακολουθούμε να αντιμετωπίζουμε ένα σημαντικό ερώτημα: πόσο πιθανό είναι να βρίσκεται μια τέτοια εκτίμηση κοντά στην πραγματική τιμή;

Για παράδειγμα, ας υποθέσουμε ότι προσπαθείτε να εκτιμήσετε το ποσοστό των ημερών που βρέχει στη Ρόδο. Προφανώς, θα πρέπει να χρησιμοποιήσετε την εκτιμήτρια

$$\frac{\text{(αριθμός ημερών που βρισκόσασταν στη Ρόδο όταν έβρεχε)}}{\text{(αριθμός ημερών που βρισκόσασταν στη Ρόδο)}}$$

Ωστόσο, αν είχατε πάει στη Ρόδο μόνο μία ημέρα και έτυχε να βρέχει εκείνη την ημέρα, η εκτίμησή σας ότι στη Ρόδο βρέχει καθημερινά δεν είναι πιθανό να είναι και τόσο ακριβής. Αν, όμως, είχατε ζήσει στη Ρόδο δέκα χρόνια, θα μπορούσατε να εκτιμήσετε το ποσοστό των βροχερών ημερών με πολύ μεγαλύτερη ακρίβεια.

Η τιμή ενός στατιστικού στοιχείου που χρησιμοποιείται ως εκτιμήτρια εξαρτάται από τις τιμές μιας ομάδας τυχαίων μεταβλητών, κάτι που σημαίνει ότι και η ίδια η εκτιμήτρια είναι μια τυχαία μεταβλητή. Θα ήταν πολύ χρήσιμο αν μπορούσαμε να βρούμε με τι μοιάζει η κατανομή της εκτιμήτριας. Για παράδειγμα, ας υποθέσουμε ότι χρησιμοποιούμε το μέσο όρο του δείγματος $\overline{x}$ για να εκτιμήσουμε την τιμή του μέσου μιας κανονικής τυχαίας μεταβλητής:

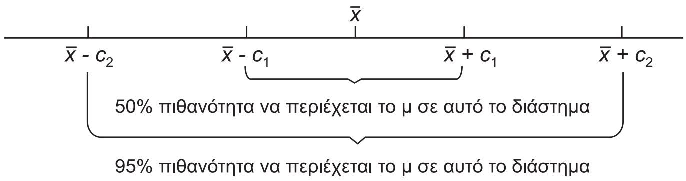
$$\overline{x} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

Καθένα από αυτά τα $X$ έχει κανονική κατανομή˙ έτσι, με βάση την ιδιότητα της πρόσθεσης των κανονικών τυχαίων μεταβλητών, το $\overline{x}$ θα πρέπει να έχει και αυτό κανονική κατανομή (το $\overline{x}$ υπολογίζεται με την άθροιση μιας ομάδας κανονικών τυχαίων μεταβλητών). Έχουμε ήδη βρει (δείτε τις σελίδες 153-154) ότι $E(\overline{x}) = \mu$ και $\mathrm{Var}(\overline{x}) = \sigma^2/n$.

## ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΣΤΗΜΑΤΩΝ ΕΜΠΙΣΤΟΣΥΝΗΣ ΤΟΥ ΜΕΣΟΥ ΟΤΑΝ ΕΙΝΑΙ ΓΝΩΣΤΗ Η ΔΙΑΚΥΜΑΝΣΗ

Τώρα που γνωρίζουμε την κατανομή του $\overline{x}$, μπορούμε να είμαστε πιο ακριβείς σχετικά με την ποιότητα της εκτίμησής μας. Ξέρουμε ότι η πραγματική τιμή του μ είναι πιθανό να βρίσκεται κοντά στο $\overline{x}$ — όμως πόσο κοντά; Είναι πιθανό να βρίσκεται το $\overline{x}$ 1 μονάδα μακριά από το μ ή μήπως είναι πιθανό να βρίσκεται 50 μονάδες μακριά του; Θα θέλαμε, λοιπόν, να γνωρίζουμε την πιθανότητα να είναι η απόσταση του $\overline{x}$ από το μ μικρότερη από μια συγκεκριμένη τιμή $c$. Με άλλα λόγια, θέλουμε να ξέρουμε την πιθανότητα να κυμαίνεται η πραγματική τιμή του μ από $(\overline{x} - c)$ μέχρι $(\overline{x} + c)$.

Είναι προφανές ότι η πιθανότητα εξαρτάται σε μεγάλο βαθμό από το $c$ που θα επιλέξουμε (δείτε την Εικόνα 11-1). Αν επιλέξουμε κάποια μεγάλη τιμή για το $c$, θα μπορούμε να είμαστε σχεδόν απόλυτα βέβαιοι ότι η τιμή του μ θα βρίσκεται στο διάστημα. Θα μπορούσαμε, για παράδειγμα, να ορίσουμε το $c$ έτσι ώστε να είναι ίσο με το άπειρο. Τότε, η πιθανότητα να βρίσκεται το μ στο διάστημα θα είναι 100%, μιας και είναι προφανές ότι το μ θα βρίσκεται ανάμεσα στο ($\overline{x}$ − άπειρο) και το ($\overline{x}$ + άπειρο). Ωστόσο, ένα τόσο μεγάλο διάστημα δεν είναι και πολύ χρήσιμο. Αν κάνουμε το διάστημα "στενότερο" επιλέγοντας κάποια μικρότερη τιμή για το $c$, θα μπορούμε να είμαστε περισσότερο ακριβείς για την πραγματική τιμή του μ. Από την άλλη μεριά, όταν στενεύουμε το διάστημα αυξάνεται η πιθανότητα να μην περιέχεται το μ στο διάστημα.




**Εικόνα 11-1**

Η συνηθισμένη διαδικασία που ακολουθείται στη στατιστική είναι η εξής: Καταρχάς, επιλέγουμε την πιθανότητα που θέλουμε — με άλλα λόγια, ορίζουμε προκαταβολικά την πιθανότητα να βρίσκεται το μ στο διάστημα. Συχνά, αυτή η πιθανότητα ορίζεται ίση με 95%. Στη συνέχεια υπολογίζουμε το εύρος που πρέπει να έχει το διάστημα ώστε να υπάρχει πιθανότητα 95% να περιέχει την πραγματική τιμή. Αυτού του είδους το διάστημα ονομάζεται **διάστημα εμπιστοσύνης** (confidence interval), και η πιθανότητα 95% είναι το **επίπεδο εμπιστοσύνης** (confidence level).

Τώρα θα πρέπει να υπολογίσουμε την τιμή του $c$ που ικανοποιεί την εξίσωση

$$\Pr(\bar{x} - c < \mu < \bar{x} + c) = 0{,}95$$

ή

$$\Pr(-c < \bar{x} - \mu < c) = 0{,}95$$

Από τη στιγμή που θα γνωρίζουμε την τιμή του $c$, θα ξέρουμε και πόσο μεγάλο θα πρέπει να είναι το διάστημα εμπιστοσύνης. Αυτό σημαίνει ότι δουλειά μας είναι να υπολογίσουμε την τιμή του $c$. Ας ορίσουμε, λοιπόν, μια νέα τυχαία μεταβλητή που θα την ονομάσουμε $Z$:

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\,\frac{\bar{x} - \mu}{\sigma}$$

Αν χρησιμοποιήσουμε τις ιδιότητες που ορίσαμε για τις κανονικές τυχαίες μεταβλητές, ξέρουμε ότι η $Z$ έχει τυπική κανονική κατανομή (με μέσο 0 και διακύμανση 1· δείτε το Κεφάλαιο 8). Μπορούμε, λοιπόν, να ξαναγράψουμε την εξίσωση έτσι:

$$\Pr\left(\frac{-c\sqrt{n}}{\sigma} < \sqrt{n}\,\frac{\bar{x} - \mu}{\sigma} < \frac{c\sqrt{n}}{\sigma}\right) = 0{,}95$$

$$\Pr\left(\frac{-c\sqrt{n}}{\sigma} < Z < \frac{c\sqrt{n}}{\sigma}\right) = 0{,}95$$

Αυτό το πρόβλημα απαιτεί τη χρήση του πίνακα των πιθανοτήτων της τυπικής κανονικής κατανομής. Ας ορίσουμε, λοιπόν, το $a$ ως εξής: $a = \sqrt{n}/\sigma$. Τότε

$$\Pr(-a < Z < a) = 0{,}95$$

Τώρα θα πρέπει να ψάξουμε στον Πίνακα Α3-2 για να βρούμε μια τιμή του $a$ που ικανοποιεί την εξίσωση. Ψάχνοντας, λοιπόν, μπορούμε να βρούμε ότι, σε αυτή την περίπτωση, η κατάλληλη τιμή για το $a$ είναι 1,96.

Τώρα μπορούμε να υπολογίσουμε την τιμή του $c$:

$$c = \frac{1{,}96\sigma}{\sqrt{n}}$$

Έτσι, λοιπόν, ξέρουμε πόσο μεγάλο πρέπει να είναι το διάστημα εμπιστοσύνης. Υπάρχει πιθανότητα 95% ότι το διάστημα από το $\overline{x} - 1{,}96\sigma/\sqrt{n}$ μέχρι το $\overline{x}$ + $1{,}96\sigma/\sqrt{n}$ θα περιέχει την πραγματική τιμή του μ.

Δύο χαρακτηριστικά αυτού του αποτελέσματος βασίζονται στην απλή λογική. Πρώτον, όσο μεγαλύτερο είναι το σ, τόσο πλατύτερο είναι το διάστημα εμπιστοσύνης (δηλαδή, υπάρχει μεγαλύτερη αβεβαιότητα). Αν η διακύμανση κάθε μεμονωμένης παρατήρησης είναι μεγαλύτερη, είναι δυσκολότερο να εντοπίσουμε την πραγματική τιμή του μ. Δεύτερον, όσο μεγαλύτερο είναι το $n$, τόσο στενότερο είναι το διάστημα εμπιστοσύνης. Αυτό σημαίνει ότι, καθώς συλλέγουμε όλο και περισσότερες παρατηρήσεις, μπορούμε να προβλέψουμε την πραγματική τιμή του μ με μεγαλύτερη ακρίβεια.

Αν θέλουμε, μπορούμε να είμαστε ακόμη περισσότερο επιφυλακτικοί. Ας υποθέσουμε ότι θέλουμε να είμαστε 99% βέβαιοι ότι το διάστημα εμπιστοσύνης θα περιέχει την πραγματική τιμή του μ. Σε αυτή την περίπτωση, θα πρέπει να συμβιβαστούμε με ένα πλατύτερο και λιγότερο ακριβές διάστημα. Από την άλλη μεριά, αν δεν θέλαμε να είμαστε και τόσο προσεκτικοί, θα μπορούσαμε να υπολογίσουμε ένα μικρότερο διάστημα εμπιστοσύνης που θα είχε μικρότερη πιθανότητα να περιέχει την πραγματική τιμή.

Τα διαστήματα εμπιστοσύνης που μας απασχολούν εδώ (και που θα μας απασχολήσουν και αργότερα) είναι συμμετρικά με τον εξής τρόπο: η πιθανότητα ότι το αριστερό άκρο του διαστήματος είναι μικρότερο από την παράμετρο (το μ εδώ) είναι ίδια με την πιθανότητα ότι το δεξιό άκρο του διαστήματος είναι μεγαλύτερο από την παράμετρο. Θεωρητικά, θα μπορούσαμε να ορίσουμε και ασύμμετρα διαστήματα εμπιστοσύνης, αλλά τα συμμετρικά είναι τα στενότερα που μπορούμε να ορίσουμε και, γι' αυτόν το λόγο, τα ακριβέστερα.

**Γενική διαδικασία υπολογισμού του διαστήματος εμπιστοσύνης του μέσου όταν έχουμε *n* παρατηρήσεις μιας κανονικής κατανομής με γνωστή τυπική απόκλιση σ**

1. Αποφασίζουμε τι είδους διάστημα εμπιστοσύνης θέλουμε. Αν θέλουμε να είμαστε περισσότερο προσεκτικοί, πρέπει να επιλέξουμε κάποιο υψηλό επίπεδο εμπιστοσύνης (ένα από τα πιο συνηθισμένα είναι για πιθανότητα 0,95).
2. Ψάχνουμε για την τιμή του $a$ στον Πίνακα Α3-2. Αν ονομάσουμε ΕΕ το επίπεδο εμπιστοσύνης, τότε

$$\Pr(-a < Z < a) = \text{ΕΕ}$$

3. Υπολογίσουμε τα $\overline{x}$ και $a\sigma/\sqrt{n}$.

   Το επίπεδο εμπιστοσύνης είναι από $\overline{x} - a\sigma/\sqrt{n}$ έως $\overline{x} + a\sigma/\sqrt{n}$.

## ΤΙ ΠΡΕΠΕΙ ΝΑ ΘΥΜΑΣΤΕ

1. Όταν υπολογίζουμε μια εκτίμηση για μια άγνωστη παράμετρο του πληθυσμού, χρειάζεται να ξέρουμε πόσο ακριβής είναι πιθανό να είναι αυτή η εκτίμηση.
2. Είναι πολύ χρήσιμος ο υπολογισμός ενός *διαστήματος εμπιστοσύνης* — δηλαδή, ενός διαστήματος ορισμένου με τέτοιο τρόπο ώστε να υπάρχει μια σταθερή πιθανότητα ότι το διάστημα θα περιέχει την άγνωστη τιμή της παραμέτρου του πληθυσμού.
3. Αυτή η σταθερή πιθανότητα είναι γνωστή ως *επίπεδο εμπιστοσύνης* και συχνά ορίζεται ίση με 95%.
4. Ένα στενό διάστημα εμπιστοσύνης είναι προτιμότερο γιατί σημαίνει ότι έχουμε τη δυνατότητα να κάνουμε ακριβέστερη εκτίμηση για την πραγματική τιμή της παραμέτρου.
5. Γενικά, το διάστημα εμπιστοσύνης γίνεται στενότερο όσο αυξάνεται το πλήθος των παρατηρήσεων.

## ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΣΤΗΜΑΤΩΝ ΕΜΠΙΣΤΟΣΥΝΗΣ ΜΕ ΧΡΗΣΗ ΤΗΣ ΚΑΤΑΝΟΜΗΣ *t*

Ο τρόπος υπολογισμού των διαστημάτων εμπιστοσύνης που περιγράψαμε κρύβει μια σημαντική δυσκολία. Πολύ συχνά δεν ξέρουμε την πραγματική τιμή του $\sigma^2$. Σε μια πρώτη προσέγγιση θα μπορούσαμε να εκτιμήσουμε την τιμή του $\sigma^2$ χρησιμοποιώντας τη διακύμανση του δείγματος. Αποδεικνύεται ότι, αν το μέγεθος του δείγματος ($n$) είναι αρκετά μεγάλο (για παράδειγμα, αν $n > 30$), μπορούμε να χρησιμοποιήσουμε τον τύπο υπολογισμού του διαστήματος εμπιστοσύνης που περιγράψαμε στην προηγούμενη ενότητα χρησιμοποιώντας αντί για το $\sigma^2$ τη διακύμανση του δείγματος $s_1^2$. Ωστόσο, για τα μικρά δείγματα πρέπει να αναπτύξουμε μια άλλη μέθοδο.

Θυμηθείτε ότι ο αρχικός υπολογισμός του διαστήματος εμπιστοσύνης βασίστηκε στο γεγονός ότι η μεταβλητή

$$Z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$$

έχει τυπική κανονική κατανομή. Ας ορίσουμε μια νέα τυχαία μεταβλητή που θα ονομάσουμε $T$:

$$T = \frac{\sqrt{n}(\bar{x} - \mu)}{s_2}$$

Προσέξτε ότι η $T$ είναι ίδια με τη $Z$ με τη διαφορά ότι, αντί για την άγνωστη τιμή του σ χρησιμοποιείται η γνωστή τιμή $s_2 = \sqrt{s_2^2}$. Θα περιμέναμε, λοιπόν, ότι η κατανομή της $T$ θα μοιάζει πολύ με την τυπική κανονική κατανομή.

Στο Κεφάλαιο 8 διαπιστώσαμε ότι η κατανομή $t$ μοιάζει με την κανονική κατανομή. Έτσι, η μεταβλητή που ονομάσαμε $T$ θα έχει κατανομή $t$ με $n - 1$ βαθμούς ελευθερίας.

Τώρα μπορούμε να υπολογίσουμε το εύρος του διαστήματος εμπιστοσύνης. Το μόνο που χρειάζεται είναι να βρούμε την τιμή του $c$ που ικανοποιεί την εξίσωση

$$\Pr(\bar{x} - c < \mu < \bar{x} + c) = 0{,}95$$

Χρησιμοποιώντας τον ορισμό της μεταβλητής $T$ μπορούμε να ξαναγράψουμε την εξίσωση ως εξής:

$$\Pr\left(\frac{-c\sqrt{n}}{s_2} < T < \frac{c\sqrt{n}}{s_2}\right) = 0{,}95$$

Ορίζουμε το $a$ ίσο με $c\sqrt{n}/s_2$.

Τώρα απομένει να ψάξουμε σε έναν πίνακα της κατανομής $t$ για να βρούμε μια τιμή του $a$ τέτοια ώστε

$$\Pr(-a < T < a) = 0{,}95$$

Για παράδειγμα, αν $n - 1 = 8$, μπορούμε να βρούμε από τον Πίνακα Α3-5 ότι $a = 2{,}306$. Από τη στιγμή που θα βρούμε το $a$, μπορούμε να υπολογίσουμε το $c$ με τη βοήθεια του τύπου

$$c = s_2 \frac{a}{\sqrt{n}}$$

Κατά συνέπεια, το διάστημα εμπιστοσύνης του 95% για το μ είναι από $\bar{x} - s_2 a/\sqrt{n}$ έως $\bar{x} + s_2 a/\sqrt{n}$.

*Παράδειγμα: Εύρεση του διαστήματος εμπιστοσύνης με τη χρήση της κατανομής t*

ΠΡΟΒΛΗΜΑ Μας έχει δοθεί ένας κατάλογος 750 πόλεων σε ολόκληρο τον κόσμο και θέλουμε να εκτιμήσουμε το μέσο πληθυσμό αυτών των πόλεων.

ΛΥΣΗ Επιλέγουμε ένα τυχαίο δείγμα 20 από αυτές τις πόλεις. Για να επιλέξουμε τις πόλεις αρκεί να χρησιμοποιήσουμε ένα πρόγραμμα υπολογιστή παραγωγής τυχαίων αριθμών. Ακολουθεί ένα παράδειγμα 20 τυχαία επιλεγμένων πόλεων:

| Πληθυσμός | Πόλη |
|---:|---|
| 566.000 | Βίλνιους (Λιθουανία) |
| 266.500 | Βισμπάντεν (Γερμανία) |
| 8.243.400 | Βομβάη (Ινδία) |
| 923.300 | Γκουιάνια (Βραζιλία) |
| 4.884.200 | Δελχί (Ινδία) |
| 112.300 | Θάντερ Μπέι (Καναδάς) |
| 1.144.000 | Ιμπαντάν (Νιγηρία) |
| 160.800 | Κάντιθ (Ισπανία) |
| 1.056.400 | Κιτακιόσου (Ιαπωνία) |
| 1.279.200 | Κουριτίμπα (Βραζιλία) |
| 346.900 | Μαϊάμι (ΗΠΑ) |
| 473.800 | Μπανγκούι (Κεντροαφρικανική Δημοκρατία) |
| 550.000 | Μπουκαραμάνγκα (Κολομβία) |
| 285.000 | Νουάκτσοτ (Μαυριτανία) |
| 1.287.600 | Ρεσίφ (Βραζιλία) |
| 818.300 | Σακάι (Ιαπωνία) |
| 484.700 | Σκάρμπορο (Καναδάς) |
| 348.000 | Τουλούζ (Γαλλία) |
| 336.500 | Τούσον (ΗΠΑ) |
| 1.170.000 | Τσενγκ Τσου (Κίνα) |

Για να υπολογίσουμε ένα διάστημα εμπιστοσύνης 95% για το μέσο, πρέπει να υπολογίσουμε πρώτα τα $\bar{x}$ = 1.236.845 και $s_2$ = 1.923.221,09740. Αφού $n$ = 20, πρέπει να ψάξουμε στον Πίνακα Α3-5 για 20 – 1 = 19 βαθμούς ελευθερίας, οπότε και θα βρούμε $a$ = 2,093. Κατά συνέπεια, το διάστημα εμπιστοσύνης είναι 1.236.845 ± 2,093 × 1.939.221,09740/$\sqrt{20}$, δηλαδή από 329.272 έως 2.144.417.

Αυτό το διάστημα είναι τόσο πλατύ ώστε να μην είναι και πολύ χρήσιμο. Το πρόβλημα είναι ότι οι πληθυσμοί των πόλεων έχουν πολύ μεγάλη διακύμανση — υπάρχουν μερικές πολύ μεγάλες πόλεις και πολλές μικρότερες. Οι πληθυσμοί των πόλεων δεν ακολουθούν την κανονική κατανομή αλλά γνωρίζουμε από το κεντρικό οριακό θεώρημα ότι, όπως και να είναι, το $\overline{x}$ θα έχει κανονική κατανομή. Όπως μπορείτε να μαντέψετε, ο μόνος τρόπος για να πετύχουμε στενότερο διάστημα εμπιστοσύνης είναι να επιλέξουμε μεγαλύτερο δείγμα.

Σημείωση: είναι σωστό να πούμε ότι "υπάρχει 95% πιθανότητα ότι το διάστημα $\overline{x} - s_2\, a/\sqrt{n}$ έως $\overline{x} + s_2\, a/\sqrt{n}$ θα περιέχει την πραγματική τιμή του μ". Τα δύο άκρα του διαστήματος είναι τυχαίες μεταβλητές. Πάντως, μετά την εκτέλεση του υπολογισμού δεν είναι σωστό να πούμε ότι "υπάρχει 95% πιθανότητα ότι το διάστημα από 329.272 έως 2.144.417 θα περιέχει την πραγματική τιμή του μ". Επειδή το μ δεν είναι τυχαία μεταβλητή, δεν έχει νόημα να μιλάμε για την πιθανότητα να παίρνει τιμές από ένα συγκεκριμένο εύρος.

**Διαδικασία υπολογισμού διαστημάτων εμπιστοσύνης με τη χρήση της κατανομής *t* όταν έχουμε *n* παρατηρήσεις της τυχαίας μεταβλητής *X***

(Μπορεί να χρησιμοποιηθεί σε μικρά δείγματα όταν η διακύμανση σ είναι άγνωστη.)

1. Αποφασίζουμε ποιο θα είναι το επίπεδο εμπιστοσύνης (ΕΕ). (Συχνά χρησιμοποιείται το διάστημα 95%.)
2. Υπολογίζουμε το $\overline{x}$:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

3. Υπολογίζουμε το $s_2$:

$$s_2 = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1}} =$$

$$= \sqrt{\frac{n}{n-1}\left(\overline{x^2} - \overline{x}^2\right)}$$

4. Ψάχνουμε για την τιμή του $a$ στον Πίνακα Α3-5.

$$\Pr(-a < T < a) = \text{ΕΕ}$$

όπου η μεταβλητή $T$ έχει κατανομή $t$ με $n - 1$ βαθμούς ελευθερίας. (Σημειώστε ότι, αν το $n$ είναι μεγαλύτερο από 30, η κατανομή $t$ είναι σχεδόν πανομοιότυπη με την τυπική κανονική κατανομή.)

Το διάστημα εμπιστοσύνης του μ είναι από $\overline{x} - s_2\, a/\sqrt{n}$ έως $\overline{x} + s_2\, a/\sqrt{n}$.

## ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΔΙΑΣΤΗΜΑΤΟΣ ΕΜΠΙΣΤΟΣΥΝΗΣ ΤΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

Υποθέστε ότι έχετε ένα δείγμα που λήφθηκε από έναν κατά προσέγγιση κανονικό πληθυσμό του οποίου τη διακύμανση θα θέλατε να εκτιμήσετε. Ξέρουμε πώς υπολογίζεται η διακύμανση ενός δείγματος $s_1^2$ και θα θέλαμε τώρα να βρούμε το διάστημα εμπιστοσύνης της διακύμανσης του πληθυσμού $\sigma^2$. Έχουμε δει ότι η μεταβλητή

$$Y^2 = \frac{ns_1^2}{\sigma^2}$$

είναι μια τυχαία μεταβλητή $\chi^2$ με $n - 1$ βαθμούς ελευθερίας. Για δύο οποιουσδήποτε θετικούς αριθμούς $a$ και $b$ ξέρουμε ότι

$$\Pr(a < Y^2 < b) = \Pr\left(a < \frac{ns_1^2}{\sigma^2} < b\right) = \Pr\left(\frac{ns_1^2}{b} < \sigma^2 < \frac{ns_1^2}{a}\right)$$

Για παράδειγμα, ακολουθούν τα ύψη (σε εκατοστά του μέτρου) 25 μαθητών ενός δημοτικού σχολείου που επιλέχθηκαν στην τύχη:

135, 139, 128, 143, 122, 123, 142, 135, 140, 141, 115, 133, 128,
137, 142, 128, 135, 142, 129, 133, 141, 137, 125, 127, 138

Μπορούμε να υπολογίσουμε ότι $s_1^2$ είναι 54, 17. Θα ορίσουμε ένα διάστημα εμπιστοσύνης 90% για τη διακύμανση. Πρέπει, λοιπόν, να βρούμε δύο αριθμούς $a$ και $b$ τέτοιους ώστε

$$\Pr(Y^2 < a) = 0{,}05 \quad \text{και} \quad \Pr(Y^2 < b) = 0{,}95$$

Αν ψάξουμε στον Πίνακα Α3-3 της κατανομής $\chi^2$ με $25 - 1 = 24$ βαθμούς ελευθερίας, θα βρούμε $a = 13{,}85$ και $b = 36{,}4$. Κατά συνέπεια, το διάστημα εμπιστοσύνης είναι από $ns_1^2/b = 37{,}20$ έως $ns_1^2/a = 97{,}78$.

**Διαδικασία υπολογισμού του διαστήματος εμπιστοσύνης της διακύμανσης**

1. Αποφασίζουμε ποιο θα είναι το επίπεδο εμπιστοσύνης (ΕΕ).
2. Υπολογίζουμε τα $\overline{x}$, $\overline{x^2}$, και $s_1^2 = \overline{x^2} - \overline{x}^2$.
3. Ανατρέχουμε στον Πίνακα Α3-3 για να βρούμε τα $a$ και $b$ που θα είναι τέτοια ώστε, αν η $Y^2$ είναι μια τυχαία μεταβλητή με κατανομή $\chi^2$ και $n - 1$ βαθμούς ελευθερίας, θα ισχύει

$$\Pr(Y^2 < a) = \frac{1 - \text{ΕΕ}}{2} \quad \text{και} \quad \Pr(Y^2 < b) = \frac{1 + \text{ΕΕ}}{2}$$

Για παράδειγμα, αν ΕΕ = 0,90, τότε $\Pr(Y^2 < a) = 0{,}05$ και $\Pr(Y^2 < b) = 0{,}95$.

Το διάστημα εμπιστοσύνης της διακύμανσης $\sigma^2$ είναι από $ns_1^2/b$ έως $ns_1^2/a$.

## ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΔΙΑΣΤΗΜΑΤΟΣ ΕΜΠΙΣΤΟΣΥΝΗΣ ΤΗΣ ΔΙΑΦΟΡΑΣ ΔΥΟ ΜΕΣΩΝ

Πολύ συχνά θα έρχεστε αντιμέτωποι με μια κατάσταση στην οποία θα θέλετε να συγκρίνετε δύο πληθυσμούς σε ό,τι αφορά κάποια τυχαία μεταβλητή. Να μερικά παραδείγματα:

- το μέσο εισόδημα σε δύο διαφορετικές πόλεις
- τα έσοδα από τις πωλήσεις δύο εταιρειών
- το πλήθος των αναγνωστών δύο εφημερίδων

Σε όλα τα παραπάνω παραδείγματα θα είχε μεγάλο ενδιαφέρον η εκτίμηση της διαφοράς των δύο μέσων ($\mu_a - \mu_b$) των δύο τυχαίων μεταβλητών $X_a$ (με μέσο $\mu_a$ και διακύμανση $\sigma_a^2$) και $X_b$ (με μέσο $\mu_a$ και διακύμανση $\sigma_b^2$). Αυτό γίνεται με τον υπολογισμό της διαφοράς των μέσων των δειγμάτων ($\overline{x}_a - \overline{x}_b$). Όπως και πριν, μας απασχολεί το θέμα της ακρίβειας. Πόσο κοντά θα βρίσκεται η τιμή $\overline{x}_a - \overline{x}_b$ στην άγνωστη τιμή $\mu_a - \mu_b$; Για άλλη μια φορά θα πρέπει να καταφύγουμε στα διαστήματα εμπιστοσύνης.

Ας υποθέσουμε ότι η $X_a$ έχει μέσο δείγματος $\overline{x}_a$ και μέγεθος δείγματος $n_a$, και ότι, με παρόμοιο τρόπο, η $X_b$ έχει μέσο δείγματος $\overline{x}_b$ και μέγεθος δείγματος $n_b$. Αν οι $X_a$ και $X_b$ είναι κανονικές τυχαίες μεταβλητές, τότε το ίδιο ισχύει και για τη διαφορά $\overline{x}_a - \overline{x}_b$, η οποία έχει μέσο $\mu_a - \mu_b$ και διακύμανση ($\sigma_a^2/n_a + \sigma_b^2/n_b$). Όπως και νωρίτερα, ορίζουμε μια νέα τυχαία μεταβλητή $Z$:

$$Z = \frac{(\overline{x}_a - \overline{x}_b) - (\mu_a - \mu_b)}{\sqrt{\sigma_a^2/n_a + \sigma_b^2/n_b}}$$

η οποία έχει τυπική κανονική κατανομή. Στη συνέχεια, ανατρέχουμε στον Πίνακα Α3-2 για να βρούμε μια τιμή του $a$ τέτοια ώστε

$$\Pr(-a < Z < a) = \text{ΕΕ}$$

όπου ΕΕ είναι το επίπεδο εμπιστοσύνης. Κατόπιν, υπολογίζουμε το $c$:

$$c = a\sqrt{\sigma_a^2/n_a + \sigma_b^2/n_b}$$

Έτσι, το διάστημα εμπιστοσύνης της διαφοράς $\mu_a - \mu_b$ είναι:

$$(\overline{x}_a - \overline{x}_b) \pm c$$

Για παράδειγμα, ακολουθεί ένα σύνολο εσόδων από πωλήσεις (σε χιλιάδες φύλλα) μιας εφημερίδας σε δύο γειτονικές πόλεις για μια περίοδο λίγων ημερών:

Πόλη Α: 25, 13, 14, 19, 23, 30, 35, 29, 28, 17, 17, 16, 13, 18, 20

Πόλη Β: 10, 12, 15, 13, 7, 6, 11, 5, 9, 14, 15, 18, 17, 16,
12, 12, 10, 11, 13, 14

Υποθέστε ότι $\sigma_a^2 = 40$ και $\sigma_b^2 = 14$. Μπορούμε να υπολογίσουμε ότι $\overline{x}_a = 21{,}13$, $n_a = 15$, $\overline{x}_b = 12{,}00$, $n_b = 20$, και $\overline{x}_a - \overline{x}_b = 9{,}13$. Αν χρησιμοποιήσουμε ένα διάστημα εμπιστοσύνης 99%, βρίσκουμε ότι $a = 2{,}58$, $c = 2{,}58\sqrt{40/15 + 14/20} = 4{,}73$, και το διάστημα εμπιστοσύνης είναι από 4,40 έως 13,86.

Έτσι, μπορούμε να συνοψίσουμε τα παραπάνω στην ακόλουθη διαδικασία (παρόμοια με αυτή που χρησιμοποιούμε για τον υπολογισμό του διαστήματος εμπιστοσύνης του μέσου μιας μεταβλητής):

**Διαδικασία υπολογισμού του διαστήματος εμπιστοσύνης της διαφοράς δύο μέσων (όταν τα $\sigma_a$ και $\sigma_b$ είναι γνωστά)**

1. Αποφασίζουμε ποιο διάστημα εμπιστοσύνης θέλουμε να έχουμε (για παράδειγμα, 95%).
2. Αναζητούμε την τιμή του $a$ στον Πίνακα Α3-2.
3. Υπολογίζουμε τη διαφορά $\overline{x}_a - \overline{x}_b$.
4. Υπολογίζουμε το $c = a\sqrt{\sigma_a^2/n_a + \sigma_b^2/n_b}$.

   Το διάστημα εμπιστοσύνης είναι $\overline{x}_a - \overline{x}_b \pm c$.

Και πάλι, αν τα μεγέθη των δειγμάτων είναι αρκετά μεγάλα, στη θέση των διακυμάνσεων των πληθυσμών $\sigma_a^2$ και $\sigma_b^2$ (όταν είναι άγνωστες) μπορούν να χρησιμοποιηθούν οι διακυμάνσεις των δειγμάτων. Όταν τα δείγματα είναι μικρά καταφεύγουμε στην κατανομή *t*. Πάντως, σε αυτό το σημείο θα πρέπει να κάνουμε ακόμη δύο υποθέσεις — ότι $\sigma_a^2 = \sigma_b^2$ και ότι τα δείγματα επιλέγονται ανεξάρτητα το ένα από το άλλο. (Όταν τα δείγματα είναι μικρά, είναι πολύ επικίνδυνο να χάσουμε την τυχαιότητα που είναι ουσιώδης στην καλή στατιστική.)

Θυμηθείτε ότι δεν υποθέτουμε ότι γνωρίζουμε τις διακυμάνσεις $\sigma_a^2$ και $\sigma_b^2$. Απλώς υποθέτουμε ότι είναι ίσες. Αν τις ξέρουμε, καλώς. Αν όχι, μπορούμε να τις εκτιμήσουμε κατά προσέγγιση με τη *συγχωνευμένη εκτίμηση* (pooled estimate) $s_p^2$:

$$s_p^2 = \frac{(n_a - 1)\, s_a^2 + (n_b - 1)\, s_b^2}{(n_a - 1) + (n_b - 1)}$$

Θα χρησιμοποιούμε το $s_a^2$ για να συμβολίζουμε τη διακύμανση (2η παραλλαγή) του δείγματος $a$:

$$s_a^2 = \frac{(\overline{x_a^2} - \overline{x}_a^2)\, n_a}{n_a - 1}$$

Παρόμοια, με το $s_b^2$ θα συμβολίζουμε τη διακύμανση (2η παραλλαγή) του δείγματος $b$. Επίσης, η $s_p^2$ είναι ο σταθμικός μέσος όρος των $s_a^2$ και $s_b^2$, και είναι τέτοια ώστε, αν το $n_a$ είναι πολύ μεγαλύτερο από το $n_b$, η $s_p^2$ να είναι πιο κοντά στην $s_a^2$ απ' ό,τι στην $s_b^2$, και αντίστροφα.

Τώρα, στον τελευταίο τύπο της προηγούμενης σελίδας για τα μεγάλα δείγματα, αν $\sigma_a^2 = \sigma_b^2$, έχουμε

$$Z = \frac{(\bar{x}_a - \bar{x}_b) - (\mu_a - \mu_b)}{\sqrt{\sigma_a^2/n_a + \sigma_b^2/n_b}} = \frac{(\bar{x}_a - \bar{x}_b) - (\mu_a - \mu_b)}{\sqrt{\sigma_a^2(1/n_a + 1/n_b)}}$$

Στη συνέχεια, ορίζουμε μια νέα μεταβλητή $T$ με κατανομή $t$ αντικαθιστώντας τη διακύμανση $\sigma_a^2$ με τη συγχωνευμένη εκτιμήτρια $s_p^2$:

$$T = \frac{(\bar{x}_a - \bar{x}_b) - (\mu_a - \mu_b)}{\sqrt{\sigma_p^2(1/n_a + 1/n_b)}}$$

Αυτή η μεταβλητή $T$ έχει $(n_a - 1) + (n_b - 1) = n_a + n_b - 2$ βαθμούς ελευθερίας. Αν γνωρίζουμε την τιμή της $\sigma_a^2$, αντικαθιστούμε με αυτή την $s_p^2$ (δεν υπάρχει λόγος να κάνουμε εκτιμήσεις όταν έχουμε την πραγματική τιμή).

Κατόπιν, αναζητούμε στον Πίνακα Α3-5 την τιμή του $a$ έτσι ώστε

$$\Pr(-a < T < a) = \text{ΕΕ}$$

όπου ΕΕ είναι το επίπεδο εμπιστοσύνης, και η μεταβλητή $T$ έχει κατανομή $t$ με $n_a + n_b - 2$ βαθμούς ελευθερίας. Στη συνέχεια, υπολογίζουμε το $c$:

$$c = a\sqrt{s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}$$

Κατά συνέπεια, το διάστημα εμπιστοσύνης της διαφοράς $\mu_a - \mu_b$ είναι

$$(\bar{x}_a - \bar{x}_b) \pm c$$

Για παράδειγμα, ακολουθούν οι πωλήσεις (σε χιλιάδες κομμάτια) δύο διαφορετικών τύπων χαρταετών την Καθαρά Δευτέρα σε τυχαία επιλεγμένες χρονιές:

Τύπος Α: 15, 20, 33, 27

Τύπος Β: 23, 42, 39

Υποθέτουμε ότι $\sigma_a^2 = \sigma_b^2$. Μπορούμε να υπολογίσουμε ότι $\bar{x}_a = 23{,}75$, $\bar{x}_b = 34{,}67$, $\bar{x}_a - \bar{x}_b = -10{,}92$, $s_a^2 = 62{,}25$, $s_b^2 = 104{,}3$, και $s_p^2 = 79{,}08$. Αν αποφασίσουμε να χρησιμοποιήσουμε ένα διάστημα εμπιστοσύνης 95% για $n_a + n_b - 2 = 5$ βαθμούς ελευθερίας, βρίσκουμε ότι $a = 2{,}571$, $c = 2{,}571\sqrt{79{,}08\,(1/4 + 1/3)} = 17{,}46$, και το διάστημα εμπιστοσύνης είναι από $-28{,}38$ έως $6{,}54$.

**Διαδικασία υπολογισμού του διαστήματος εμπιστοσύνης της διαφοράς των μέσων δύο πληθυσμών όταν έχουμε μικρά δείγματα και ίσες διακυμάνσεις πληθυσμών**

1. Αποφασίζουμε ποιο θα είναι το διάστημα εμπιστοσύνης (για παράδειγμα, 95%).
2. Υπολογίζουμε τη διαφορά $\overline{x}_a - \overline{x}_b$.
3. Υπολογίζουμε τη συγχωνευμένη εκτιμήτρια $s_p^2$:

$$s_p^2 = \frac{(n_a - 1)\, s_a^2 + (n_b - 1)\, s_b^2}{n_a + n_b - 2}$$

4. Αναζητούμε την τιμή του $a$ στον Πίνακα Α3-5 για $n_a + n_b - 2$ *βαθμούς ελευθερίας.*
5. Υπολογίζουμε το $c$:

$$c = a\sqrt{s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}$$

Το διάστημα εμπιστοσύνης είναι $(\overline{x}_a - \overline{x}_b) \pm c$.

# ΓΝΩΣΗ ΤΩΝ ΕΝΝΟΙΩΝ

## ΓΝΩΡΙΖΕΤΕ ΤΑ ΒΑΣΙΚΑ;

Ελέγξτε το βαθμό στον οποίο κατανοήσατε τις έννοιες του Κεφαλαίου 11 απαντώντας στις παρακάτω ερωτήσεις:

1. Γιατί να μην υπολογίσουμε ένα διάστημα εμπιστοσύνης με πολύ μεγάλο επίπεδο εμπιστοσύνης — ας πούμε 99,99%;
2. Γιατί ένα διάστημα εμπιστοσύνης είναι πιο αξιόλογο από μια μεμονωμένη σημειακή εκτίμηση;
3. Αναφέρετε τρεις τρόπους με τους οποίους μπορούμε να στενέψουμε ένα διάστημα εμπιστοσύνης.
4. Γιατί να πρέπει να υπολογίσουμε ένα διάστημα εμπιστοσύνης χρησιμοποιώντας την κατανομή *t* αντί για την κανονική κατανομή;
5. Αν διπλασιάσουμε το πλήθος των παρατηρήσεων θα μπορέσουμε να μειώσουμε το εύρος ενός διαστήματος εμπιστοσύνης στο μισό;

## ΟΡΟΙ ΓΙΑ ΜΕΛΕΤΗ

διάστημα εμπιστοσύνης
επίπεδο εμπιστοσύνης

# ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

## ΠΡΟΒΛΗΜΑΤΑ ΥΠΟΛΟΓΙΣΜΟΥ

*Κάθε μια από τις παρακάτω λίστες αριθμών αντιπροσωπεύει εβδομαδιαία ποσά καθαρών κερδών (σε εκατομμύρια δραχμές) μιας εταιρείας. Υποθέστε ότι, σε κάθε εταιρεία, τα ποσά των καθαρών κερδών έχουν κανονική κατανομή με άγνωστο μέσο και άγνωστη διακύμανση. Υπολογίστε ένα διάστημα εμπιστοσύνης 95% της πραγματικής τιμής του μέσου.*

| | | | | | |
|---|---|---|---|---|---|
| **1.** | 23 | 12 | 1 | 6 | 4 |
| | 16 | 28 | 14 | 6 | 28 |
| | 15 | 18 | 6 | 2 | 14 |
| | 19 | 11 | 15 | 20 | 20 |
| **2.** | 5 | 48 | 6 | 49 | 0 |
| | 11 | 44 | 47 | 35 | 13 |
| | 10 | 4 | 50 | 7 | 22 |
| | 1 | 40 | 4 | 4 | 26 |
| | 8 | 18 | 35 | 23 | 12 |
| **3.** | 12 | 1 | 22 | 8 | 24 |
| | 25 | 22 | 21 | 13 | 12 |
| | 18 | 8 | 0 | 16 | 20 |
| | 11 | | | | |
| **4.** | 10 | 20 | 19 | 30 | 17 |
| | 28 | 26 | 26 | 29 | 28 |
| | 6 | 4 | | | |
| **5.** | 23 | 8 | 18 | 29 | 14 |
| | 20 | 1 | 19 | 19 | 28 |
| | 18 | 16 | 30 | 4 | 19 |

*Σε κάθε μια από τις επόμενες ασκήσεις δίνεται το σύνολο μιας λίστας παρατηρήσεων, ο μέσος όρος αυτών των παρατηρήσεων, και η τυπική απόκλιση ($s_2$) αυτών των παρατηρήσεων. Υπολογίστε ένα διάστημα εμπιστοσύνης 95% και ένα διάστημα εμπιστοσύνης 99% για την άγνωστη τιμή του μέσου κάθε κατανομής.*

**6.** Σύνολο: 523,114 Μέσος όρος: 52,311 $s_2$: 8,813

**7.** Σύνολο: 1.571,322 Μέσος όρος: 104,755 $s_2$: 3,002

**8.** Σύνολο: 1.150,295 Μέσος όρος: 127,811 $s_2$: 4,611

**9.** Σύνολο: 4.528,186 Μέσος όρος: 283,012 $s_2$: 8,386

**10.** Σύνολο: 2.957,990 Μέσος όρος: 147,899 $s_2$: 1,688

# ΑΠΑΝΤΗΣΕΙΣ

## ΜΑΘΕΤΕ ΤΑ ΒΑΣΙΚΑ

1. Ένα διάστημα με τόσο μεγάλο επίπεδο εμπιστοσύνης θα είναι πάρα πολύ μεγάλο ώστε να είναι χρήσιμο.
2. Μια σημειακή εκτίμηση δεν μας δίνει καμία πληροφορία για το πόσο ακριβής είναι πιθανό να είναι αυτή η εκτίμηση.
3. Περισσότερες παρατηρήσεις, μικρότερη τιμή για τη διακύμανση του πληθυσμού, χαμηλότερο επίπεδο εμπιστοσύνης.
4. Η κατανομή *t* θα πρέπει να χρησιμοποιείται όταν η διακύμανση του πληθυσμού είναι άγνωστη και το πλήθος των παρατηρήσεων είναι μικρότερο από 30.
5. Όχι. Το πλάτος του διαστήματος εμπιστοσύνης είναι ανάλογο με την ποσότητα $1/\sqrt{n}$.

## ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

Ακολουθούν οι υπολογισμοί για την Άσκηση 1 και την Άσκηση 6. Οι ίδιες μέθοδοι θα πρέπει να χρησιμοποιηθούν και για τις Ασκήσεις 2-5 και 7-10, αντίστοιχα.

**1.** $\overline{x} = (23 + 12 + 1 + 6 + 4 + 16 + 28 + 14 + 6 + 28 + 15 + 18 + 6 + 2 + 14 + 19 + 11 + 15 + 20 + 20)/20 = 13{,}9$

$$\overline{x^2} = (23^2 + 12^2 + 1^2 + 6^2 + 4^2 + 16^2 + 28^2 + 14^2 + 6^2 + 28^2 + 15^2 + 18^2 + 6^2 + 2^2 + 14^2 + 19^2 + 11^2 + 15^2 + 20^2 + 20^2)/20 = 253{,}7$$

$$s_2 = \sqrt{253{,}7 - 13{,}9^2} \times \sqrt{19/20} = 7{,}98$$

Από τον Πίνακα A3-5 μπορούμε να βρούμε ότι, για μια κατανομή *t* με 19 βαθμούς ελευθερίας, $a = 2{,}093$. Κατά συνέπεια, το διάστημα εμπιστοσύνης 95% είναι

$$13{,}9 \pm 2{,}093 \times 7{,}98/\sqrt{20}$$ — δηλαδή, από 10,165 έως 17,635

**2.** Μέσος όρος = 20,880 $s_2 = 17{,}405$
διάστημα εμπιστοσύνης 95% = 13,695 έως 28,065

**3.** Μέσος όρος = 14,562 $s_2 = 7{,}763$
διάστημα εμπιστοσύνης 95% = 10,427 έως 18,698

**4.** Μέσος όρος = 20,250 $s_2$ = 9,255
διάστημα εμπιστοσύνης 95% = 14,369 έως 26,131

**5.** Μέσος όρος = 17,733 $s_2$ = 8,455
διάστημα εμπιστοσύνης 95% = 13,050 έως 22,416

| | Διαστήματα εμπιστοσύνης 95% | Διαστήματα εμπιστοσύνης 99% |
|---|---|---|
| **6.** | $52{,}311 \pm 2{,}262 \times 8{,}813/\sqrt{10}$ | $52{,}311 \pm 3{,}250 \times 8{,}813/\sqrt{10}$ |
| | 46,007 έως 58,615 | 43,254 έως 61,368 |
| **7.** | 103,092 έως 106,417 | 102,447 έως 107,063 |
| **8.** | 124,266 έως 131,355 | 122,654 έως 132,968 |
| **9.** | 278,544 έως 287,479 | 276,834 έως 289,190 |
| **10.** | 147,110 έως 148,689 | 146,819 έως 148,979 |