Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

# Στατιστική για Πολιτικούς Μηχανικούς

## Λυμένες ασκήσεις μέρους Β’

Κουγιουμτζής Δημήτρης
Τμήμα Πολιτικών Μηχανικών Α.Π.Θ.

Θεσσαλονίκη, Μάρτιος 2014

## Άδειες Χρήσης

Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons. Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.


## Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα. Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

ανοικτάμαθήματα
opencourses

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ
2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Μέρος Β – Στατιστική

**Άσκηση 1** *[Θέμα στις εξετάσεις Φεβρουαρίου 2002]*

**(α)** Αν δύο τυχαίες μεταβλητές $X_1$ και $X_2$ έχουν κοινή διασπορά $\sigma^2$, και $s_1^2$, $s_2^2$ είναι οι αμερόληπτες δειγματικές διασπορές των $X_1$ και $X_2$, αντίστοιχα, από δείγματα μεγέθους $n_1$ και $n_2$, δείξτε ότι η εκτιμήτρια $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ της $\sigma^2$ είναι επίσης αμερόληπτη.

**(β)** Δύο εργοστάσια Α και Β παραγωγής χάλυβα θέλουν να εκτιμήσουν την περιεκτικότητα του χάλυβα σε ραδιενέργεια και γι αυτό έκαναν τις παρακάτω μετρήσεις ραδιενέργειας (η ραδιενέργεια μετριέται σε Bq/g) σε τυχαία δοκίμια χάλυβα:

| Δοκίμια | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Α (Bq/g) | 0.37 | 0.00 | 0.54 | 0.59 | 0.16 | 0.86 | 0.86 | 0.49 | 0.60 | 0.55 | | | | | |
| Β (Bq/g) | 0.24 | 0.52 | 0.12 | 0.95 | 0.26 | 0.33 | 0.62 | 0.32 | 0.27 | 0.05 | 0.39 | 0.10 | 0.51 | 0.79 | 0.09 |

Θεωρούμε ότι η περιεκτικότητα του χάλυβα σε ραδιενέργεια ακολουθεί κανονική κατανομή και η διασπορά της ραδιενέργειας στο χάλυβα είναι ίδια για τα δύο εργοστάσια ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

**(i)** Εκτιμήστε τη μέση ραδιενέργεια στο χάλυβα για το εργοστάσιο Α και Β (σημειακή εκτίμηση και 95% διάστημα εμπιστοσύνης).

**(ii)** Το μέσο ανώτατο επιτρεπτό όριο μέσης ραδιενέργειας στο χάλυβα είναι 0.5Bq/g. Με βάση τα παραπάνω δείγματα και με εμπιστοσύνη σε επίπεδο 95% θα γινόταν αποδεκτός στην αγορά ο χάλυβας από το εργοστάσιο Α? Από το εργοστάσιο Β?

**(iii)** Ελέγξτε σε επίπεδο εμπιστοσύνης 95% αν η μέση ραδιενέργεια στο χάλυβα των δύο εργοστασίων είναι ίδια.

[Για την απάντηση σας στα ερωτήματα (ii) και (iii) μπορείτε να χρησιμοποιήσετε διάστημα εμπιστοσύνης **ή** στατιστικό έλεγχο]

**Λύση**

**(α)** Για να είναι η εκτιμήτρια $s_p^2$ της κοινής διασποράς $\sigma^2$ θα πρέπει $\mathrm{E}(s_p^2) = \sigma^2$. Έχουμε

$$\mathrm{E}(s_p^2) = \mathrm{E}\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right) = \frac{(n_1-1)\mathrm{E}(s_1^2) + (n_2-1)\mathrm{E}(s_2^2)}{n_1+n_2-2} = \frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1+n_2-2}$$
$$= \frac{(n_1-1)\sigma^2 + (n_2-1)\sigma^2}{n_1+n_2-2} = \sigma^2.$$

**(β)** Η ραδιενέργεια του χάλυβα στα δύο εργοστάσια είναι οι δύο τ.μ. $X_1$ και $X_2$ που ακολουθούν κανονική κατανομή και έχουν κοινή αλλά άγνωστη διασπορά ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). Έχουμε $n_1 = 10$, $n_2 = 15$.

**(i)** Υπολογίζουμε τις δειγματικές μέσες τιμές, διασπορές και τυπικές αποκλίσεις για τα δύο δείγματα. **(**τυπολόγιο: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ , $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$ **)**

$\bar{x}_1 = 0.502$ , $s_1^2 = 0.07377$ , $s_1 = 0.272$

$\bar{x}_2 = 0.371$, $s_2^2 = 0.06979$ , $s_2 = 0.264$.

Οι δύο δειγματικές μέσες τιμές $\bar{x}_1 = 0.502$ και $\bar{x}_2 = 0.371$ αποτελούν τις σημειακές εκτιμήσεις της μέσης ραδιενέργειας χάλυβα στα δύο εργοστάσια. Το 95% διάστημα εμπιστοσύνης (δ.ε.) της μέσης ραδιενέργειας χάλυβα στα δύο εργοστάσια δίνεται από τον τύπο για μικρό δείγμα, και τ.μ. με άγνωστη διασπορά που ακολουθεί κανονική κατανομή, κάνοντας χρήση της κρίσιμης τιμής $t_{1-\alpha/2,n-1}$ της κατανομής student, δηλαδή είναι $\bar{x} \pm t_{1-a/2,n-1} \cdot \frac{s}{\sqrt{n}}$ (τυπολόγιο).

Για το πρώτο εργοστάσιο, επίπεδο σημαντικότητας $\alpha = 0.05$, έχουμε από τον στατιστικό πίνακα για την κατανομή student $t_{0.975,9} = 2.26$ και το 95% δ.ε. είναι

$$0.502 \pm 2.26\frac{0.272}{\sqrt{10}} \longrightarrow 0.502 \pm 0.194 \longrightarrow [0.308,\ 0.696].$$

Αντίστοιχα για το δεύτερο εργοστάσιο έχουμε την κρίσιμη τιμή $t_{0.975,14} = 2.14$ και το 95% δ.ε. είναι

$$0.371 \pm 2.14\frac{0.264}{\sqrt{15}} \longrightarrow 0.371 \pm 0.146 \longrightarrow [0.225,\ 0.517].$$

Η μέση ραδιενέργεια του χάλυβα είναι πολύ πιθανό (με εμπιστοσύνη σε επίπεδο 95%) να βρίσκεται μεταξύ 0.308 Bq/g και 0.696 Bq/g για το εργοστάσιο Α και μεταξύ 0.225 Bq/g και 0.517 Bq/g για το εργοστάσιο Β.

**(ii)** <u>*Χρησιμοποιώντας διάστημα εμπιστοσύνης.*</u>: Με βάση τα παραπάνω αποτελέσματα, το μέσο ανώτατο επιτρεπτό όριο μέσης ραδιενέργειας στο χάλυβα είναι 0.5 Bq/g περιέχεται στο 95% δ.ε. της μέσης ραδιενέργειας και για τα δύο εργοστάσια, που σημαίνει ότι η μέση ραδιενέργεια του χάλυβα μπορεί να ξεπεράσει το μέσο επιτρεπτό όριο. Άρα ο χάλυβας και από τα δύο εργοστάσια δε θα γίνει αποδεκτός.

<u>*Χρησιμοποιώντας έλεγχο υπόθεσης.*</u>: Για να απαντήσουμε στο ερώτημα, μπορούμε εναλλακτικά να κάνουμε έλεγχο υπόθεσης για το αν η μέση ραδιενέργεια μπορεί να πάρει την τιμή $\mu_0 = 0.5$.

$H_0$ : $\mu = \mu_0$

$H_1$ : $\mu \neq \mu_0$ [δίπλευρος έλεγχος]

Για μικρό δείγμα και τ.μ. με άγνωστη διασπορά που ακολουθεί κανονική κατανομή, η στατιστική για τον έλεγχο αυτό είναι $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$ (δεν υπάρχει στο τυπολόγιο, προκύπτει άμεσα από το αντίστοιχο δ.ε. στο τυπολόγιο).

Η απορριπτική περιοχή $R$ για επίπεδο σημαντικότητας $\alpha = 0.05$ σχηματίζεται από την κρίσιμη τιμή $t_{1-\alpha/2,n-1}$ της κατανομής student : $R = \{t \mid |t| > t_{1-\alpha/2,n-1}\}$.

Για το εργοστάσιο Α είναι $R = \{t \mid |t| > 2.26\}$. Η δειγματική στατιστική είναι

$\tilde{t} = \dfrac{0.502 - 0.5}{0.272/\sqrt{10}} = 0.023$. Ισχύει $\tilde{t} \notin R$ και δεν απορρίπτεται η $H_0$.

Για το εργοστάσιο Α είναι $R = \{t \mid |t| > 2.14\}$. Η δειγματική στατιστική είναι

$\tilde{t} = \dfrac{0.371 - 0.5}{0.264/\sqrt{15}} = -1.04$. Ισχύει και πάλι $\tilde{t} \notin R$ και δεν απορρίπτεται η $H_0$. Άρα και για τα δύο εργοστάσια η μέση ραδιενέργεια του χάλυβα μπορεί να είναι 0.5 Bq/g, δηλαδή να ξεπεράσει το μέσο επιτρεπτό όριο και ο χάλυβας και από τα δύο εργοστάσια δε θα γίνει αποδεκτός.

Θα μπορούσαμε να κάνουμε μονόπλευρο έλεγχο, δηλαδή να εξετάσουμε για την εναλλακτική υπόθεση $H_1: \mu < \mu_0$, όποτε και θα άλλαζε η απορριπτική περιοχή, δηλαδή θα ήταν $R = \{t \mid t < -t_{1-\alpha, n-1}\}$ (θα εξετάζαμε μόνο αν η δειγματική στατιστική $\tilde{t}$ μπορεί να βρίσκεται στην αριστερή ουρά της κατανομής student). Τα συμπεράσματα θα ήταν τα ίδια αφού οι δειγματικές στατιστικές και για τα δύο δείγματα δεν είναι κοντά στην αριστερή ουρά της κατανομής student.

**(iii)** Υπολογίζουμε πρώτα τη διαφορά των δειγματικών μέσων τιμών $\bar{x}_1 - \bar{x}_2 = 0.131$ και την εκτίμηση της κοινής διασποράς

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{9 \cdot 0.07377 + 14 \cdot 0.06979}{23} = 0.07135.$$

(δεν υπάρχει στο τυπολόγιο, προκύπτει ως σταθμισμένος μέσος των δύο δειγματικών διασπορών σταθμίζοντας με τους βαθμούς ελευθερίας κάθε δείγματος (μέγεθος δείγματος - 1)

*Χρησιμοποιώντας διάστημα εμπιστοσύνης.*: Χρησιμοποιούμε το 95% διάστημα εμπιστοσύνης (δ.ε.) για τη διαφορά της μέσης ραδιενέργειας χάλυβα στα δύο εργοστάσια $\mu_1 - \mu_2$. Εδώ έχουμε ότι τα δύο δείγματα είναι μικρά και οι τ.μ. ακολουθούν κανονική κατανομή αλλά με άγνωστη και κοινή διασπορά. Γι αυτό κάνουμε χρήση του τύπου που βασίζεται στην κατανομή student και η κρίσιμη τιμή είναι $t_{0.975,23} = 2.07$. Το 95% δ.ε. είναι

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-a/2, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{(τυπολόγιο)}$$

και έχουμε

$$0.131 \pm 2.06 \cdot 0.267 \cdot \sqrt{\frac{1}{10} + \frac{1}{15}} \longrightarrow 0.131 \pm 0.226 \longrightarrow [-0.094, 0.357].$$

Παρατηρούμε ότι το δ.ε. περιέχει το 0, έστω και οριακά, δηλαδή η διαφορά $\mu_1 - \mu_2$ μπορεί να είναι και 0, άρα οι δύο μέσες ραδιενέργειες σε χάλυβα του εργοστασίου Α και Β δε φαίνεται να διαφέρουν (σε επίπεδο εμπιστοσύνης 95%).

*Χρησιμοποιώντας έλεγχο υπόθεσης.*: Ελέγχουμε την υπόθεσης η μέση ραδιενέργεια να είναι ίδια στους χάλυβες των δύο εργοστασίων.

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

$H_1: \mu_1 \neq \mu_2$ [δίπλευρος έλεγχος]

Η στατιστική για τον έλεγχο αυτόν ακολουθεί κατανομή student και είναι

$$t \equiv \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

(δεν υπάρχει στο τυπολόγιο, προκύπτει άμεσα από το αντίστοιχο δ.ε. στο τυπολόγιο)

Η απορριπτική περιοχή $R$ για επίπεδο σημαντικότητας $\alpha = 0.05$ σχηματίζεται από την κρίσιμη τιμή $t_{0.975,23} = 2.07$ της κατανομής student : $R = \{t \mid |t| > 2.07\}$. Η δειγματική στατιστική είναι

$\tilde{t} = \dfrac{0.131}{0.267\sqrt{1/10 + 1/15}} = 0.201$. Ισχύει $\tilde{t} \notin R$ και δεν απορρίπτεται η $H_0$, οριακά όμως καθώς η τιμή της δειγματικής στατιστικής $\tilde{t}$ είναι πολύ κοντά στην κρίσιμη τιμή για τη δεξιά ουρά.

Έτσι παρ' όλο που φαίνεται (με δ.ε. και έλεγχο υπόθεσης) η μέση ραδιενέργεια στο χάλυβα του εργοστασίου Α να είναι μεγαλύτερη από αυτή του εργοστασίου Β, η διαφορά αυτή δε βρέθηκε σημαντική σε επίπεδο εμπιστοσύνης 95%.

**Άσκηση 2** *[Θέμα στις εξετάσεις Φεβρουαρίου 2002]*

Στον παρακάτω πίνακα δίνεται για 10 σταθμούς ο αριθμός των ημερών σ' ένα χρόνο που η θερμοκρασία έπεσε κάτω από 0°C και το υψόμετρο τους.

| Υψόμετρο (μ) | 1000 | 1050 | 1110 | 1220 | 1320 | 1380 | 1420 | 1560 | 1670 | 1950 |
|---|---|---|---|---|---|---|---|---|---|---|
| Αριθμός ημερών | 32 | 29 | 36 | 38 | 43 | 53 | 52 | 63 | 73 | 100 |

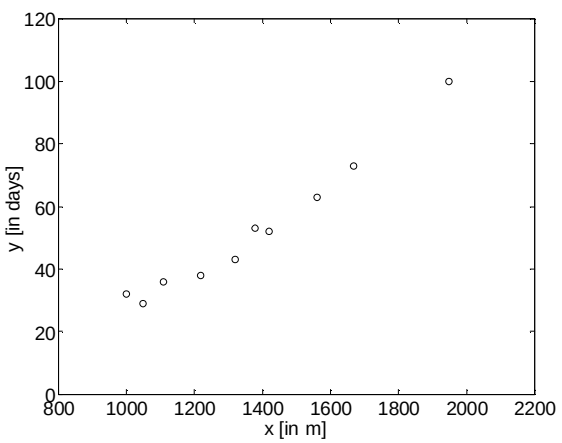
**(α)** Υποθέτουμε ότι ο αριθμός των ημερών $Y$ εξαρτάται γραμμικά από το υψόμετρο $X$ ( $E(Y \mid X=x) = \alpha+\beta x$ ). Σχηματίστε το κατάλληλο διάγραμμα διασποράς και σχολιάστε αν αυτή η υπόθεση φαίνεται σωστή με βάση το δείγμα των παρατηρήσεων του πίνακα.

**(β)** Υπολογίστε τις σημειακές εκτιμήσεις a και b των παραμέτρων α και β της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων).

**(γ)** Με βάση το δείγμα, μπορείτε να εκτιμήσετε το μέσο αριθμό ημερών το χρόνο που η θερμοκρασία πέφτει κάτω από 0°C σε υψόμετρο 1500μ; Σε υψόμετρο 2200μ;

**Λύση**

**(α)** Η ανεξάρτητη μεταβλητή $X$ είναι το υψόμετρο του σταθμού και η εξαρτημένη μεταβλητή $Y$ είναι ο αριθμός των ημερών σ' ένα χρόνο που η θερμοκρασία έπεσε κάτω από 0°C. Σχηματίζουμε το διάγραμμα διασποράς.



Από το διάγραμμα διασποράς φαίνεται να υπάρχει γραμμική θετική εξάρτηση γιατί όταν μεγαλώνει το υψόμετρο πληθαίνουν αναλογικά οι μέρες που η θερμοκρασία πέφτει κάτω από $0^oC$. Φαίνεται επίσης η εξάρτηση αυτή να είναι ισχυρή γιατί μπορούμε να καθορίσουμε με αρκετή ακρίβεια των αριθμό των ημερών ανά έτος που η θερμοκρασία πέφτει κάτω από $0^oC$ όταν γνωρίζουμε το υψόμετρο (τα σημεία βρίσκονται πολύ κοντά σε μια νοητή ευθεία).

**(β)** Έχουμε δείγμα μεγέθους $n=10$. Υπολογίζουμε τα παρακάτω:

$$\bar{x}=1368 \quad \bar{y}=51.9 \quad \sum_{i=1}^{10} x_i^2 = 19511200 \qquad \sum_{i=1}^{10} x_i \cdot y_i = 767700$$

και βρίσκουμε τη δειγματική διασπορά της $X$ καθώς και τη δειγματική συνδιασπορά των $X$ και $Y$:

$$s_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{9}\left(19511200 - 10\cdot 1368^2\right) = 88551.11 \qquad \text{(τυπολόγιο)}$$

$$s_{XY} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right) = \frac{1}{9}\left(767700 - 10\cdot 1368\cdot 51.9\right) = 6412\,. \quad \text{(τυπολόγιο)}$$

Στη συνέχεια εκτιμούμε τις παραμέτρους της ευθείας ελαχίστων τετραγώνων, δηλαδή του μοντέλου γραμμικής παλινδρόμησης:

$$b = \frac{s_{XY}}{s_X^2} = \frac{6412}{88551.11} = 0.0724 \qquad \text{(τυπολόγιο)}$$

$$a = \bar{y} - b\cdot\bar{x} = 51.9 - 0.0724\cdot 1368 = -47.16 \qquad \text{(τυπολόγιο)}$$

και η ευθεία ελαχίστων τετραγώνων είναι $y = -47.16 + 0.0724\cdot x$.

**(γ)** Κάνουμε προβλέψεις χρησιμοποιώντας την ευθεία ελαχίστων τετραγώνων για υψόμετρα μέσα στο εύρος του δείγματος από 1000μ μέχρι 1950μ.

Για υψόμετρο $x_0 = 1500$, έχουμε $y = -47.16 + 0.0724\cdot 1500 = 61.46$ και άρα περιμένουμε 62 μέρες το χρόνο να πέφτει η θερμοκρασία κάτω από $0^oC$.

Για υψόμετρο $x_0 = 2200$ δε μπορούμε να κάνουμε πρόβλεψη γιατί δεν είναι μέσα στο εύρος γνωστών υψομέτρων για τα οποία ισχύει το γραμμικό μοντέλο.

**Άσκηση 3** *[Θέμα στις εξετάσεις Φεβρουαρίου 2004]*

Δίνονται οι παρακάτω μετρήσεις

5 8 10 3 4 10 7 32 9 7

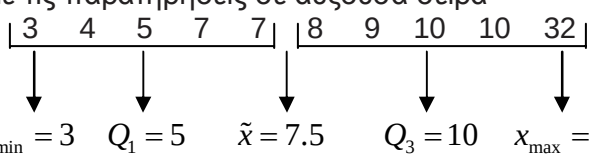
**(α)** Σχεδιάστε το θηκόγραμμα αφού εξηγήσετε πως προέκυψαν οι 5 αριθμοί που χρησιμοποιήσατε για να το σχεδιάσετε.

**(β)** Σχολιάστε αν η κατανομή της μεταβλητής στην οποία αναφέρονται οι μετρήσεις φαίνεται να είναι κανονική.

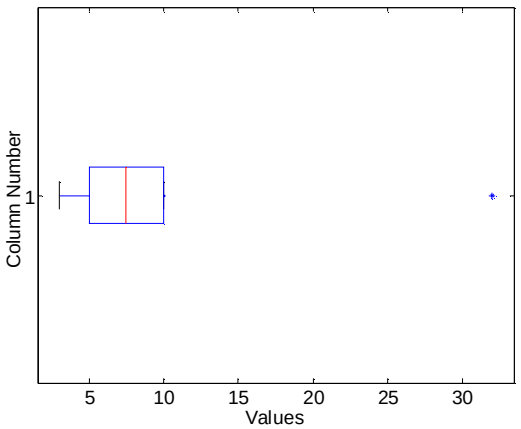
**Λύση**

**(α)** Παραθέτουμε τις παρατηρήσεις σε αύξουσα σειρά

βρίσκουμε:



και σχηματίζουμε το θηκόγραμμα, όπως στο παρακάτω σχήμα (σε κατακόρυφη θέση).




**(β)** Το δείγμα περιέχει μια απόμακρη τιμή, την τιμή 32. Η ύπαρξη μιας τόσης ακραίας τιμής σε ένα μικρό δείγμα 10 παρατηρήσεων δημιουργεί κάποια ανησυχία για το αν η κατανομή της τυχαίας μεταβλητής είναι κανονική.

**Άσκηση 4** *[Θέμα στις εξετάσεις Φεβρουαρίου 2004]*

Ένας δείκτης της κυκλοφορίας οχημάτων είναι ο αριθμός χιλιομέτρων που κάνει ένα όχημα το χρόνο. Για μια περιοχή Α συλλέξαμε ένα τυχαίο δείγμα 200 αυτοκινήτων και καταγράψαμε για κάθε αυτοκίνητο τον αριθμό χιλιομέτρων που διένυσε τον τελευταίο χρόνο. Δίνονται τα παρακάτω αποτελέσματα για το δείγμα:

$$\text{μέση τιμή } \bar{x} = 14500 \text{ km}, \qquad \text{τυπική απόκλιση } s = 4000 \text{ km}$$

**(α)** Υπολογίστε το 95% διάστημα εμπιστοσύνης για το μέσο αριθμό χιλιομέτρων που διανύει το χρόνο ένα αυτοκίνητο της περιοχής Α. Κάνετε το ίδιο για 99% επίπεδο εμπιστοσύνης και συγκρίνετε τα δύο διαστήματα εμπιστοσύνης.

**(β)** Σε ίδια μελέτη που έγινε πριν 10 χρόνια είχε βρεθεί πως η τυπική απόκλιση ήταν 3000 km. Εξετάστε με βάση το νέο δείγμα και σε επίπεδο εμπιστοσύνης 90% αν μπορούμε να δεχτούμε ότι η τυπική απόκλιση δεν άλλαξε σημαντικά [μπορείτε να χρησιμοποιήσετε διάστημα εμπιστοσύνης ή έλεγχο υπόθεσης].

**Λύση**

**(α)** Έχουμε $n = 200$, $\bar{x} = 14500$ km και $s = 4000$ km. Άρα το (1-α)% διάστημα εμπιστοσύνης (δ.ε.) για το μέσο αριθμό χιλιομέτρων που διανύει το χρόνο ένα αυτοκίνητο της περιοχής Α θα δίνεται από τον τύπο για μεγάλο δείγμα και άγνωστη διασπορά, κάνοντας χρήση της κρίσιμης τιμής $z_{1-\alpha/2}$ της τυπικής κανονικής κατανομής, δηλαδή είναι $\bar{x} \pm z_{a/2} \dfrac{s}{\sqrt{n}}$ (τυπολόγιο).

Για $\alpha = 0.05$ έχουμε από τον στατιστικό πίνακα για την τυπική κανονική κατανομή $z_{0.975} = 1.96$ και το 95% δ.ε. είναι

$$14500 \pm 1.96 \frac{4000}{\sqrt{200}} \longrightarrow 14500 \pm 554.36 \longrightarrow [13945.69, 15054.36].$$

Για $\alpha = 0.01$ έχουμε από τον στατιστικό πίνακα για την τυπική κανονική κατανομή $z_{0.995} = 2.58$ και το 99% δ.ε. είναι

$$14500 \pm 2.58 \frac{4000}{\sqrt{200}} \qquad 14500 \pm 728.55 \longrightarrow [13771.45, 15228.55].$$

Παρατηρούμε ότι το 99% δ.ε. για το μέσο αριθμό χιλιομέτρων που διανύει το χρόνο ένα αυτοκίνητο της περιοχής Α είναι μεγαλύτερο από αυτό για 95% επίπεδο εμπιστοσύνης, όπως αναμένεται αφού αυξάνουμε την εμπιστοσύνη (πιθανότητα) το διάστημα αυτό να περιέχει το πραγματικό μέσο αριθμό χιλιομέτρων.

**(β)** <u>*Χρησιμοποιώντας διάστημα εμπιστοσύνης*</u>: Βρίσκουμε πρώτα το 90% δ.ε. για τη διασπορά του αριθμού χιλιομέτρων που διανύει το χρόνο ένα αυτοκίνητο της περιοχής Α. Αυτό δίνεται κάνοντας χρήση της $\mathcal{X}^2$ κατανομής και είναι

$$\left( \frac{(n-1)s^2}{\chi^2_{n-1,1-a/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,a/2}} \right)$$ (τυπολόγιο).

Για $\alpha = 0.1$ έχουμε από τον στατιστικό πίνακα για την $\mathcal{X}^2$ κατανομή ότι η αριστερή κρίσιμη τιμή είναι $\chi^2_{0.05,199} = 167.36$ και η δεξιά $\chi^2_{0.95,199} = 232.91$ (οι τιμές αυτές δεν συμπεριλαμβάνονται στον πίνακα που έχει ως 100 βαθμούς ελευθερίας). Το 90% δ.ε. για τη διασπορά είναι

$$\left[ \frac{199 \cdot 4000^2}{232.91}, \frac{199 \cdot 4000^2}{167.36} \right] \longrightarrow$$ [13670409.6, 19024743.5].

Το αντίστοιχο δ.ε. για την τυπική απόκλιση προκύπτει παίρνοντας τη τετραγωνική ρίζα των ορίων του παραπάνω διαστήματος και άρα το 90% δ.ε. για την τυπική απόκλιση του αριθμού χιλιομέτρων που διανύει το χρόνο ένα αυτοκίνητο της περιοχής Α είναι [3697.4, 4361.7]. Το διάστημα αυτό δεν περιέχει την εμπειρική τιμή 3000 km και άρα η τυπική απόκλιση άλλαξε σημαντικά από αυτήν που είχαμε εκτιμήσει πριν 10 χρόνια.

<u>*Χρησιμοποιώντας έλεγχο υπόθεσης.*</u>: Κάνουμε έλεγχο υπόθεσης για το αν η διασπορά μπορεί να πάρει την τιμή $\sigma_0^2 = 3000^2 = 9 \cdot 10^6$.

$H_0 : \sigma^2 = \sigma_0^2$

$H_1 : \sigma^2 \neq \sigma_0^2$ [δίπλευρος έλεγχος]

Η απορριπτική περιοχή $R$ για επίπεδο σημαντικότητας $\alpha = 0.1$ σχηματίζεται από τις δύο κρίσιμες τιμές της $\mathcal{X}^2$ κατανομής : $R = \left\{ \chi^2 \mid \chi^2 < 167.36 \ \vee \ \chi^2 > 232.91 \right\}$.

Η δειγματική στατιστική είναι

$$\tilde{\chi}^2 = \frac{(n-1) \cdot s^2}{\sigma_0^2} = \frac{199 \cdot 4000^2}{3000^2} = 354 .$$

(δεν υπάρχει στο τυπολόγιο, προκύπτει άμεσα από το αντίστοιχο δ.ε. στο τυπολόγιο)

Ισχύει $\tilde{\chi}^2 \in R$ και άρα απορρίπτεται η $H_0$ και συμπεραίνουμε ότι με 90% εμπιστοσύνη (πιθανότητα) δε δεχόμαστε ότι η εμπειρική τιμή 3000 km που είχαμε εκτιμήσει πριν 10 χρόνια για την τυπική απόκλιση του αριθμού χιλιομέτρων μπορεί να ισχύει και τώρα.

**Άσκηση 5** *[Θέμα στις εξετάσεις Φεβρουαρίου 2004]*

Σε μια έρευνα στις Η.Π.Α. για την επίδραση του πληθυσμού της πόλης στη συγκέντρωση του όζοντος συγκεντρώθηκαν τα παρακάτω στοιχεία. Ο πληθυσμός των πόλεων δίνεται σε εκατομμύρια και η συγκέντρωση του όζοντος που μετρήθηκε σε κάθε πόλη δίνεται σε ppb [parts per billion] ανά ώρα.

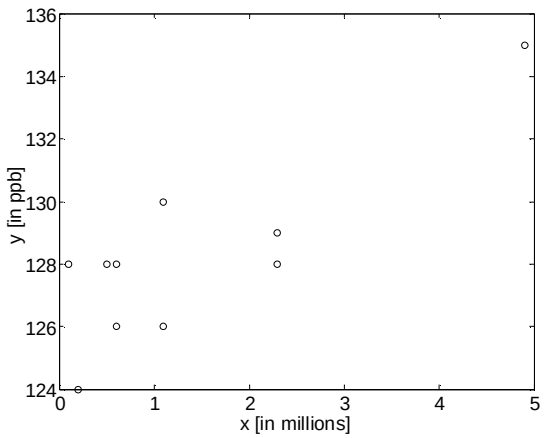
| *Πληθυσμός πόλης* | 0.1 | 0.2 | 0.5 | 0.6 | 0.6 | 1.1 | 1.1 | 2.3 | 2.3 | 4.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Συγκέντρωση όζοντος* | 128 | 124 | 128 | 126 | 128 | 126 | 130 | 128 | 129 | 135 |

**(α)** Σχηματίστε το κατάλληλο διάγραμμα διασποράς. Εκτιμείστε το συντελεστή συσχέτισης μεταξύ της συγκέντρωσης του όζοντος και του πληθυσμού της πόλης. Με βάση αυτά τα αποτελέσματα σχολιάστε αν φαίνεται να υπάρχει εξάρτηση της συγκέντρωσης του όζοντος από τον πληθυσμό της πόλης.

**(β)** Υπολογίστε τις σημειακές εκτιμήσεις *a* και *b* των παραμέτρων *α* και *β* της ευθείας παλινδρόμησης (με τη μέθοδο των ελαχίστων τετραγώνων) για το πρόβλημα της γραμμικής εξάρτησης της συγκέντρωσης του όζοντος από τον πληθυσμό της πόλης. Σχηματίστε την ευθεία ελαχίστων τετραγώνων στο διάγραμμα διασποράς που σχηματίσατε στο (α).

**Λύση**

**(α)** Σχηματίζουμε το διάγραμμα διασποράς ($X$: πληθυσμός πόλης, $Y$: συγκέντρωση όζοντος)




Από το διάγραμμα διασποράς φαίνεται να υπάρχει γραμμική θετική συσχέτιση (η αύξηση του πληθυσμού της πόλης δημιουργεί αύξηση της συγκέντρωσης όζοντος), χωρίς όμως να φαίνεται πολύ ισχυρή (δεν εξηγείται σε μεγάλο βαθμό η μεταβολή της μιας τ.μ. όταν γνωρίζουμε τη μεταβολή της άλλης, τα σημεία απλώνονται αρκετά γύρω από μια νοητή ευθεία).

Έχουμε δείγμα μεγέθους $n=10$. Υπολογίζουμε τα παρακάτω:

$$\bar{x}=1.37 \quad \bar{y}=128.2 \quad \sum_{i=1}^{10} x_i^2 = 38.03 \quad \sum_{i=1}^{10} y_i^2 = 164430 \quad \sum_{i=1}^{10} x_i \cdot y_i = 1788.2$$

και βρίσκουμε τις δειγματικές διασπορές και τυπικές αποκλίσεις των $X$ και $Y$ καθώς και τη δειγματική συνδιασπορά τους: (οι τύποι δίνονται στο τυπολόγιο)

$$s_X^2 = \frac{1}{9}\left(38.03 - 10 \cdot 1.37^2\right) = 2.140 \longrightarrow s_X = \sqrt{2.140} = 1.46$$

$$s_Y^2 = \frac{1}{9}\left(164430 - 10 \cdot 128.2^2\right) = 8.622 \longrightarrow s_Y = \sqrt{8.622} = 2.94$$

$$s_{XY} = \frac{1}{9}\left(1788.2 - 10 \cdot 1.37 \cdot 128.2\right) = 3.54 .$$

Η εκτίμηση του συντελεστή συσχέτισης μεταξύ πληθυσμού πόλης και συγκέντρωση όζοντος είναι

$$r = \frac{3.54}{1.46 \cdot 2.94} = 0.82 .$$

Η εκτίμηση του συντελεστή συσχέτισης επιβεβαιώνει ότι η συσχέτιση δεν είναι ισχυρή ($r < 0.9$).

**(β)** Η ανεξάρτητη μεταβλητή $X$ είναι ο πληθυσμός πόλης και η εξαρτημένη μεταβλητή $Y$ είναι η συγκέντρωση όζοντος. Εκτιμούμε τις παραμέτρους του μοντέλου γραμμικής παλινδρόμησης:

$$b = \frac{s_{XY}}{s_X^2} = \frac{3.54}{2.140} = 1.654 \qquad \text{(τυπολόγιο)}$$

$$a = \bar{y} - b \cdot \bar{x} = 128.2 - 1.654 \cdot 1.37 = 125.94 \qquad \text{(τυπολόγιο)}$$

και η ευθεία ελαχίστων τετραγώνων είναι $y = 125.94 + 1.654 \cdot x$ .

Για να σχηματίσουμε την ευθεία υπολογίζουμε δύο σημεία που ανήκουν σε αυτήν (καλύτερα για τη μικρότερη και μεγαλύτερη τιμή της $X$ στο δείγμα), π.χ.

$$x = 0.1 \longrightarrow y = 125.94 + 1.654 \cdot 0.1 = 126.10$$

$$x = 4.9 \longrightarrow y = 125.94 + 1.654 \cdot 4.9 = 134.04$$

και χαράζουμε το ευθύγραμμο τμήμα που περνά από αυτά τα δύο σημεία και προεκτείνεται μόνο για το εύρος των γνωστών τιμών του πληθυσμού πόλης $X$.