Πανεπιστήμιο Αιγαίου

Πιθανότητες και Στατιστική

Ελισάβετ Κωνσταντίνου

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 Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- 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 & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιγραφική Στατιστική

- Σε πρακτικές εφαρμογές η π.π. ή σ.π. της τ.μ. που μας ενδιαφέρει είναι άγνωστη.
  - Διεξαγωγή πειράματος
  - Συλλογή δεδομένων (τιμών της τ.μ.) – (τυχαίο) δείγμα από ένα πληθυσμό
  - Στατιστική επαγωγή (statistical inference) – εξαγωγή ιδιοτήτων της άγνωστης π.π.
- Ως πληθυσμό θεωρούμε γενικά το σύνολο των δυνατών παρατηρήσεων (τιμών της τ.μ.).

Παραδείγματα

- Αιτίες πτώσης του δικτύου τα τελευταία 3 χρόνια:
  - (a) Πρόβλημα γραμμών: 15
  - (b) H/W failure: 15
  - (c) S/W malfunction: 5
  - (d) Ανθρώπινο λάθος: 1
  - (e) Πτώση τάσης: 12
- Το σύνολο των 48 παρατηρήσεων:

  a a c a e a a e b b b b b c a a b e e b a e a c
  e c e a a e b e d c b a a a b e b e b b b b e a

Παράδειγμα

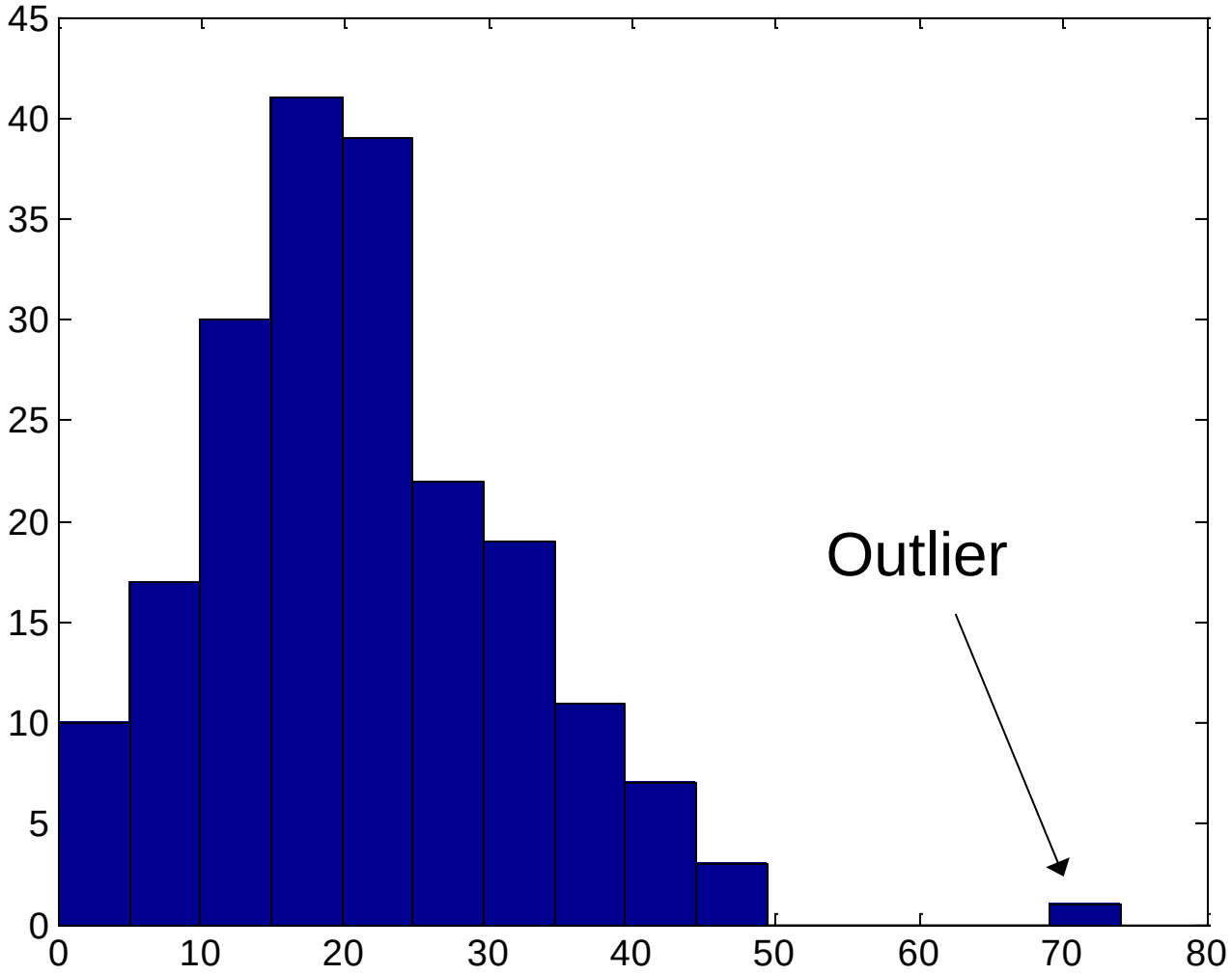
- Παρατηρήσεις του χρόνου απόκρισης Τ μιας βάσης δεδομένων:

  10.21 22.28 20.15 43.84 21.82 2.87 17.60 10.18 9.60 9.03 10.00 3.64 31.90 27.47 18.56 27.91 24.61 21.55 10.80 23.54 21.37 4.88 (...)

- Ερωτήσεις:
  - Ποιά είναι η π.π. της Τ;
  - E{T}, Var{T};
  - P(T > 30 sec);

Παράσταση δεδομένων

- Bar charts
- Pie charts
- Histograms



Δειγματικές στατιστικές συναρτήσεις (sample statistics)

- Δειγματικός μέσος (sample mean): $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Sample trimmed mean: αφαίρεση ακραίων τιμών
- Κεντρική τιμή δείγματος (sample median): (n+1)-στή τιμή ή μ.ο. της n-στής και (n+1)-στής
- Ποσοστημόρια (sample percentiles)
- Επικρατούσα τιμή (sample mode)
- Δειγματική διασπορά (sample variance):

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

Εκτίμηση παραμέτρων μιας τ.μ.

- Σημειακή εκτίμηση (point estimate) $\hat{\theta}$ μιας άγνωστης παραμέτρου θ μιας κατανομής: μια δειγματική στατιστική συνάρτηση που προσεγγίζει την θ.
- Παράδειγμα: Πιθανότητα βλάβης του δικτύου από πτώση τάσης $\hat{p}_e = 12 / 48 = 0.25$
- Παράδειγμα: Μέση τιμή και διασπορά του χρόνου απόκρισης T της βάσης δεδομένων:

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Εκτίμηση παραμέτρων

- Αμερόληπτος εκτιμητής (unbiased): $E\{\hat{\theta}\} = \theta$

  Απόκλιση της εκτίμησης (bias) = $E\{\hat{\theta}\} - \theta$
- Αν $X \sim B(n, p)$, τότε το $\hat{p} = X / n$ είναι αμερόληπτος εκτιμητής του p.
- Αν $X_1, ..., X_n$ είναι παρατηρήσεις μιας τ.μ. με μέση τιμή μ και διασπορά σ^2, τότε ο δειγματικός μέσος και η δειγματική διασπορά είναι αμερόληπτοι εκτιμητές των μ και σ^2.

Εκτίμηση παραμέτρων

- Εκτιμητής ελάχιστης διασποράς (minimum variance estimator): έχει την ελάχιστη $\mathrm{Var}\{\hat{\theta}\}$.
- Παράδειγμα: $X_1 \sim N(\theta, 2.97)$, $X_2 \sim N(\theta, 1.62)$

  $\hat{\theta}_1 = X_1 : \mathrm{Var}\{\hat{\theta}_1\} = 2.97$

  $\hat{\theta}_2 = X_2 : \mathrm{Var}\{\hat{\theta}_2\} = 1.62$

  $\hat{\theta}_3 = (X_1 + X_2) / 2 : \mathrm{Var}\{\hat{\theta}_3\} = (\mathrm{Var}\{\hat{\theta}_1\} + \mathrm{Var}\{\hat{\theta}_2\}) / 4 = 1.15$

  $\hat{\theta}_4 = aX_1 + (1–a)X_2 : \mathrm{Var}\{\hat{\theta}_4\} = a^2\mathrm{Var}\{\hat{\theta}_1\} + (1–a)^2\mathrm{Var}\{\hat{\theta}_2\}$

  Ελάχιστο για $a = 0.35$:

  $\hat{\theta}_4 = 0.35X_1 + 0.65X_2$, $\mathrm{Var}\{\hat{\theta}_4\} = 1.05$

  Βέλτιστος **γραμμικός** εκτιμητής ελάχιστης διασποράς

Εκτίμηση παραμέτρων

- Σφάλμα εκτίμησης $\hat{\theta} - \theta$: πάλι τυχαία μεταβλητή
- Μέσο τετραγωνικό σφάλμα εκτίμησης:

  $MSE(\theta) = E\{(\hat{\theta} - \theta)^2\} = E\{(\hat{\theta} - E\{\hat{\theta}\})^2\} + (E\{\hat{\theta}\} - \theta)^2$

  $= Var\{\hat{\theta}\} + bias^2$
- Εκτιμητής ελάχιστου μέσου τετραγωνικού σφάλματος (minimum mean-squared-error estimator): έχει το ελάχιστο $MSE(\hat{\theta})$.
- Συνεπής (consistent) εκτιμητής: για μεγάλο μέγεθος του δείγματος συγκλίνει στην πραγματική τιμή της θ.

$$\lim_{n \to \infty} \hat{\theta} = \theta$$

Σφάλμα της εκτίμησης

- Εκτίμηση της p μιας διωνυμικής X ~ B(n, p): $\hat{p} = X / n$

  Για μεγάλο n η κατανομή του $\hat{p}$ προσεγγίζει την κανονική $p \sim N(p, p(1-p) / n)$

- Τυπικό σφάλμα εκτίμησης (standard error):

$$s.e.(\hat{p}) = \sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n}$$

- Εκτίμηση της μέσης τιμής μ μιας τ.μ. από n ανεξάρτητες παρατηρήσεις $X_1, ..., X_n$: $\hat{\mu} = \bar{X}$.

  Για μεγάλο n, $\hat{\mu} \sim N(\mu, \sigma^2 / n)$

$$s.e.(\hat{\mu}) = \sigma / \sqrt{n} \approx s / \sqrt{n}$$

Παράδειγμα

- Εκτίμηση της πιθανότητας βλάβης του δικτύου από πτώση τάσης $\hat{p}_e = 12 / 48 = 0.25$

  Το σφάλμα εκτίμησης $\hat{p}_e - p_e \sim N(0, p_e(1-p_e) / n)$

  Τυπικό σφάλμα $s.e.(\hat{p}_e) \approx 0.0625$, δηλ. η p_e είναι μεταξύ 0.1875 και 0.3125 ($\pm 1\sigma$) με πιθανότητα $\approx 68.3\%$.

  $P(0.2 \le p_e \le 0.3) \approx \Phi(0.8) - \Phi(-8) = 0.5763$
- Αν θέλουμε $P(0.2 \le p_e \le 0.3) = 0.9$, χρειαζόμαστε μεγαλύτερο δείγμα:

  $2\Phi(0.05 / \sigma) - 1 = 0.9 \Rightarrow \Phi(0.05 / \sigma) = 0.95$
  $\Rightarrow 0.05 / \sigma = 1.645 \Rightarrow \sigma = 0.3034 \Rightarrow n = 203$

Παράδειγμα

- Ο χρόνος X μεταξύ αφίξεων πλοίων σε ένα λιμάνι έχει εκθετική κατανομή με άγνωστο λ.
- Εχουμε τις μετρήσεις X_1, ..., X_{25} (σε λεπτά):

  8 31 4 9 40 31 10 9 20 11 16 63 72
  23 87 19 8 48 67 10 10 83 82 33 11
- $\hat{\mu} = \bar{X}$: Μπορούμε να χρησιμοποιήσουμε $\hat{\lambda} = 1 / \bar{X} = 0.0311$
- $P(X > 60) \approx \exp(-60 \times 0.0311) = 0.1552$