\`Θεωρία Πιθανοτήτων και Στατιστική
Θέματα εξετάσεων και ενδεικτικές λύσεις

Θέμα 1
Γνωρίζουμε ότι 5% των ασθενών σε μία κλινική παρουσιάζουν λοίμωξη κάποιου είδους. Επίσης είναι γνωστό ότι το 40% των ασθενών που έχουν λοίμωξη έχουν μειωμένα λευκά αιμοσφαίρια ενώ το 1% των ασθενών που δεν έχουν λοίμωξη έχουν μειωμένα λευκά αιμοσφαίρια. Να υπολογισθούν οι πιθανότητες ένας ασθενής στην κλινική:
(α) να έχει μειωμένα λευκά αιμοσφαίρια.
(β) να έχει λοίμωξη δεδομένου ότι έχει μειωμένα λευκά αιμοσφαίρια.
(γ) να έχει λοίμωξη και μειωμένα λευκά αιμοσφαίρια,

Βαθμολογία: (α) 0,8μ (β) 0,9μ (γ) 0,8μ

Ενδεικτική λύση

A = {ο ασθενής έχει λοίμωξη}
B = {ο ασθενής έχει μειωμένα λευκά αιμοσφαίρια}

Από τα δεδομένα έχουμε: $P(A) = 0{,}05$, $P(B \mid A) = 0{,}4$, $P(B \mid A') = 0{,}01$.

(α) $P(B) = P(B \mid A)\, P(A) + P(B \mid A')\, P(A') = 0{,}4 \cdot 0{,}05 + 0{,}01 \cdot 0{,}95 = 0{,}0295 = 2{,}95\%$

(β) $P(A \mid B) = P(B \mid A) \cdot P(A) / P(B) = 0{,}4 \cdot 0{,}05 / 0{,}0295 = 0{,}678 = 67{,}8\%$.

(γ) $P(A\ B) = P(B \mid A) \cdot P(A) = 0{,}4 \cdot 0{,}05 = 0{,}02 = 2\%$.

Θέμα 2
Είναι γνωστό ότι το μέσο πλήθος αριθμητικών σφαλμάτων σε μία φορολογική δήλωση είναι 1,2 και πως η εμφάνιση ενός σφάλματος δεν συσχετίζεται με την εμφάνιση ενός άλλου.
(α) Να δείξετε ότι η πιθανότητα μία δήλωση να έχει τουλάχιστον 2 αριθμητικά λάθη είναι ίση με 0,337.

(β) Στην εφορεία της Ξάνθης υποβάλλονται κάθε χρόνο 27.000 φορολογικές δηλώσεις. Να βρείτε το αναμενόμενο πλήθος και την τυπική απόκλιση του πλήθους των δηλώσεων με τουλάχιστον 2 αριθμητικά λάθη.

(γ) Κάθε εφοριακός αναλαμβάνει να ελέγξει το πολύ 2.000 δηλώσεις. Να βρείτε την πιθανότητα μεταξύ 2.000 φορολογικών δηλώσεων να υπάρξουν περισσότερες από 700 δηλώσεις με τουλάχιστον 2 αριθμητικά λάθη.

Βαθμολογία: (α) 0,7μ (β) 0,7μ (γ) 1,1μ

Ενδεικτική λύση
(α) Έστω X = {πλήθος σφαλμάτων σε μία φορολογική δήλωση}. Η εμφάνιση ενός σφάλματος δεν συσχετίζεται με την εμφάνιση ενός άλλου, άρα X ~ Poisson(1,2).

Υπολογίζουμε:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) = 1 - e^{-1,2} - 1,2e^{-1,2} = 0,337.$$

(β)
Αν X_1 = {το πλήθος των δηλώσεων με τουλάχιστον 2 αριθμητικά λάθη στις 27.000} τότε

$$X_1 \sim B(27.000,\ 0,337)$$

οπότε $EX = n\cdot p = 27.000\cdot 0,337 = 9.099$ και $VarX = n\cdot p\cdot q = 27.000\cdot 0,337\cdot 0,663 = 6.032,6$, άρα η τυπική απόκλιση είναι $6.032,6^{0,5} = 77,7$ δηλώσεις.

(γ)Έστω τώρα
Y = {πλήθος δηλώσεων με τουλάχιστον 2 αριθμητικά λάθη στις 2.000 δηλώσεις}.

Είναι

$$Y \sim B(2000,\ 0.337) \approx N(674,\ 446.9) \text{ (είναι } np = 674 > 5,\ npq = 446,9 > 5)$$

Αναζητούμε την $P_B(Y > 700) = P_N(Y > 699,5) = P_N(Z > (699,5 - 674)/446,9^{0,5}) = P_N(Z > 1,206) = 1 - \Phi(1,206) = 1 - 0,897 = 0,103 = 10,3\%$.

Θέμα 3
Κατά τη διάρκεια του τρέχοντος εξαμήνου 137 φοιτητές συμμετείχαν τουλάχιστον μία φορά στις ασκήσεις εμπέδωσης που διεξαγόταν στο τέλος κάθε μαθήματος. Το δείγμα των 137 φοιτητών θεωρείται αντιπροσωπευτικό όλου του πληθυσμού των φοιτητών.

Στον πίνακα παρουσιάζεται η συμμετοχή ανά φύλο κωδικοποιημένη σε δύο κατηγορίες (1 έως 5 συμμετοχές και 6 έως 9 συμμετοχές).

Συμμετοχές	**1 έως 5**	**6 έως 9**
Κορίτσια	13	30
Αγόρια	36	58

(α) Να αποδείξετε ότι αν η συμμετοχή ενός φοιτητή σε λιγότερες ή περισσότερες από 5 ασκήσεις είναι ανεξάρτητη από το φύλο του τότε το αναμενόμενο πλήθος κοριτσιών με 6 έως 9 συμμετοχές είναι ίσο με $\frac{43 \cdot 88}{137}$.

Ο διδάσκων ισχυρίζεται πως το γεγονός της συμμετοχής ενός φοιτητή σε λιγότερες ή περισσότερες των 5 ασκήσεων εξαρτάται από το φύλο του. Εσείς καλείστε να ελέγξετε τον ισχυρισμό με μία κατάλληλη στατιστική δοκιμασία.

(β) Να αναφέρετε τη δοκιμασία και να καταγράψετε την μηδενική και την εναλλακτική της υπόθεση.
(γ) Να βρείτε αν ο ισχυρισμός του διδάσκοντος γίνεται αποδεκτός (α = 0,05).

Βαθμολογία: (α) 0,5μ (β) 0,2μ (γ) 1,8μ

Ενδεικτική λύση
(α) Αν Κ = {ο φοιτητής είναι Κορίτσι}, Σ = {ο φοιτητής συμμετείχε 6 έως 9 φορές} και η συμμετοχή ενός φοιτητή σε λιγότερες ή περισσότερες από 5 ασκήσεις είναι ανεξάρτητη από το φύλο του τότε

$$P(Κ \cap Σ) = P(Κ) \cdot P(Σ) = 43 / 137 \cdot 88 / 137 = 43 \cdot 88 / 137^2.$$

Το αναμενόμενο πλήθος φοιτητών που θα είναι κορίτσια και θα συμμετέχουν 6 έως 9 φορές είναι:

$$137 \cdot P(Κ \cap Σ) = 137 \cdot 43 \cdot 88 / 137^2 = 43 \cdot 88 / 137.$$

(β) Θα εφαρμοστεί η δοκιμασία ανεξαρτησίας χι-τετράγωνο, για να ελέγξουμε την υπόθεση
H_0: Το φύλο είναι στοχαστικά ανεξάρτητο με την κατηγορία πλήθους συμμετοχών έναντι της
H_1: Όχι η H_0.

(γ) Ο πίνακας ενημερωμένος με τις αναμενόμενες συχνότητες είναι ο εξής:

Συμμετοχές	**1 έως 5**	**6 έως 9**	**Σύνολο**
Κορίτσια	13 / 15,4	30 / 27,6	**43**
Αγόρια	36 / 33,6	58 / 60,4	**94**
Σύνολο	**49**	**88**	**137**

Παρατηρούμε ότι όλες οι αναμενόμενες συχνότητες είναι μεγαλύτερες του 5, άρα η δοκιμασία χ^2 είναι αξιόπιστη.

Υπολογίζουμε την τιμή του στατιστικού χ^2_0 να είναι:

$$\chi^2_0 = \frac{(13-15{,}4)^2}{15{,}4} + \frac{(30-27{,}6)^2}{27{,}6} + \frac{(36-33{,}6)^2}{33{,}6} + \frac{(58-60{,}4)^2}{60{,}4} = 0{,}835$$

Αυτό ακολουθεί την κατανομή $\chi^2((2-1)(2-1)) = \chi^2(1)$ και βρίσκουμε $p = P(\chi^2 > \chi^2_0) = P(\chi^2 > 0{,}835) = 0{,}361 > \alpha = 0{,}05$, άρα δεν απορρίπτεται η H_0, δηλαδή το φύλο είναι στοχαστικά ανεξάρτητο με την συμμετοχή σε λιγότερες ή περισσότερες από 5 φορές ($\chi^2(1) = 0{,}835$, $p = 0{,}361$).

Θέμα 4. Κατά τη διάρκεια του τρέχοντος εξαμήνου 137 φοιτητές συμμετείχαν τουλάχιστον μία φορά στις ασκήσεις εμπέδωσης που διεξαγόταν στο τέλος κάθε μαθήματος. Το δείγμα των 137 φοιτητών θεωρείται αντιπροσωπευτικό όλου του πληθυσμού των φοιτητών.

Το μέσο εξάμηνο φοίτησης για τα 43 κορίτσια ήταν 4,1 (SD = 1,3) ενώ για τα 94 αγόρια ήταν 4,6 (SD = 1,7). Ο διδάσκων ισχυρίζεται πως τα αγόρια που συμμετέχουν στη διαδικασία των ασκήσεων βρίσκονται σε σημαντικά μεγαλύτερο εξάμηνο σπουδών από ότι τα κορίτσια. Εσείς καλείστε να ελέγξετε τον ισχυρισμό με μία κατάλληλη στατιστική δοκιμασία.

(α) Να αναφέρετε τη δοκιμασία και να καταγράψετε την μηδενική και την εναλλακτική της υπόθεση.

(β) Να βρείτε αν ο ισχυρισμός του διδάσκοντος γίνεται αποδεκτός (α = 0,05).

(γ) Να βρείτε ένα 95% διάστημα εμπιστοσύνης για την πραγματική διαφορά $\mu_A - \mu_K$, των μέσων τιμών στα εξάμηνα σπουδών μεταξύ αγοριών και κοριτσιών.

Βαθμολογία: (α) 0,2μ (β) 1,8μ (γ) 0,5μ

Ενδεικτική λύση

(α) Θα εφαρμοστεί η δοκιμασία t – test για δύο ανεξάρτητα δείγματα, για να ελέγξουμε την υπόθεση H_0: $\mu_A \leq \mu_K$, έναντι της H_1: $\mu_A > \mu_K$ (μονόπλευρος έλεγχος).

(β) Η τυπική απόκλιση της διαφοράς των μέσων τιμών είναι

$$s_p = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\left(\frac{1}{94} + \frac{1}{43}\right)\frac{(94 - 1)1{,}7^2 + (43 - 1)1{,}3^2}{94 + 43 - 2}} = 0{,}292$$

Θα εφαρμόσουμε τη δοκιμασία t – test για δύο ανεξάρτητα δείγματα. Η τιμή του στατιστικού t_0 είναι

$$t_0 = \frac{\overline{x}_A - \overline{x}_K}{s_p} = \frac{4{,}6 - 4{,}1}{0{,}292} = 1{,}712 \quad \text{και} \quad t = \frac{\overline{X}_A - \overline{X}_K}{s_p} \sim t(135) = N(0\,,\,1)$$

Μονόπλευρος έλεγχος H_1: $\mu_A > \mu_K$: p = P(t > t_0) = P(t > 1,712) = 1 – Φ(1,712) = 1 – 0,959 = 0,041 < α = 0,05, άρα η στατιστική υπόθεση H_0: $\mu_1 \leq \mu_2$, **απορρίπτεται** έναντι της ερευνητικής υπόθεσης: H_1: $\mu_A > \mu_K$.

Ιδιαίτερα, συμπεραίνουμε ότι μεταξύ των φοιτητών που συμμετέχουν τουλάχιστον μία φορά στις ασκήσεις, τα αγόρια βρίσκονται σε σημαντικά μεγαλύτερο εξάμηνο σπουδών από ότι τα κορίτσια (t(135) = 1,712, p = 0,041).

Παρατήρηση
Είναι αξιοσημείωτο πως στην περίπτωση του δίπλευρου ελέγχου θα ήταν p = 0,082 > 0,05 και η μηδενική υπόθεση δεν θα απορριπτόταν.

(γ)

Αν $\overline{X}_A \sim N(\mu_A, \frac{\sigma^2}{n_1})$ και $\overline{X}_K \sim N(\mu_K, \frac{\sigma^2}{n_2})$ τότε $\overline{X}_A - \overline{X}_K \sim N(\mu_A - \mu_K, \sigma_p^2)$

και $\frac{\overline{X}_A - \overline{X}_K - (\mu_A - \mu_K)}{s_p} \sim t(n_1 + n_2 - 2)$ Για $n_1 = 43$ και $n_2 = 94$, είναι

$\frac{\overline{X}_A - \overline{X}_K - (\mu_A - \mu_K)}{s_p} \sim t(135) \simeq N(0, 1)$ άρα

$$P(-1,96 \le \frac{\overline{X}_A - \overline{X}_K - (\mu_A - \mu_K)}{s_p} \le 1,96) = 0,95 \quad ή$$

$$P(\overline{X}_A - \overline{X}_K - 1,96 \cdot s_p \le \mu_A - \mu_K \le \overline{X}_A - \overline{X}_K + 1,96 \cdot s_p) = 0,95$$

Συμπεραίνουμε, ότι ένα 95% διάστημα εμπιστοσύνης για την άγνωστη διαφορά είναι το

(4,6 – 4,1 – 1,96·0,292, 4,6 – 4,1 + 1,96·0,292) = (-0,072, 1,072)