Θεωρία Πιθανοτήτων και Στατιστική
Θέματα εξετάσεων Σεπτεμβρίου 2024

Θέμα 1

Βρέθηκε ότι από τα 3000 μηνύματα που έφθασαν μια χρονική περίοδο σε έναν mail server, τα 2000 είναι Spam και τα 1000 δεν είναι Spam. Βρέθηκε επίσης ότι η λέξη «Rolex» εμφανίστηκε σε 250 από τα 2000 μηνύματα που είναι Spam και σε 5 από τα 1000 μηνύματα που δεν είναι Spam. Επιλέγουμε στην τύχη ένα μήνυμα από τα 3000 που έχουν φτάσει στον mail server.

(α) Να βρεθεί η πιθανότητα το μήνυμα αυτό να είναι spam δεδομένου ότι περιέχει τη λέξη «Rolex»

(β) Να βρεθεί η πιθανότητα το μήνυμα να περιέχει τη λέξη “Rolex” δεδομένου ότι δεν είναι spam.

(γ) Να βρεθεί η πιθανότητα το μήνυμα να περιέχει τη λέξη “Rolex” και να είναι spam.

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

Θέμα 2

Σε μια βιομηχανία υπάρχουν συνεργεία που εργάζονται ανεξάρτητα το ένα από το άλλο και κατασκευάζουν μεταλλικά εξαρτήματα. Το πλήθος των εξαρτημάτων που κατασκευάζονται από όλα τα συνεργεία κάθε ημέρα δεν είναι το ίδιο, ωστόσο γνωρίζουμε πως το μέσο πλήθος ανά ημέρα είναι 7.

(α) Να δείξετε ότι η πιθανότητα σε μία ημέρα να κατασκευαστούν λιγότερα από 3 εξαρτήματα είναι ίση με 0,0296.

Η πιθανότητα για κάθε τέτοιο εξάρτημα να αντέξει σε μία ορισμένη διαδικασία καταπόνησης είναι ίση με 0,8. Επιλέγουμε τυχαία 9 εξαρτήματα και τα υποβάλλουμε στη συγκεκριμένη καταπόνηση. Ποια είναι η πιθανότητα να αντέξουν:

β) το πολύ 7 εξαρτήματα,

γ) λιγότερα από 6 και τουλάχιστον 4 εξαρτήματα.

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

Θέμα 3

Σε δημοσκόπηση 300 φοιτητών (100 άνδρες και 200 γυναίκες) καταγράφηκε το φύλο κάθε φοιτητή και το χρώμα που προτιμά μεταξύ των χρωμάτων Κόκκινο, Κίτρινο, Μπλε. Τα αποτελέσματα παρουσιάζονται στον πίνακα που ακολουθεί:

Χρώμα	Κόκκινο	Κίτρινο	Μπλε
Άνδρες	21	34	45
Γυναίκες	66	63	71

Ο δημοσκόπος ισχυρίζεται ότι η επιλογή του χρώματος συσχετίζεται σημαντικά με το φύλο του φοιτητή. 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 καταγράψετε τις αναμενόμενες θεωρητικές συχνότητες που θα παρατηρούσαμε για κάθε συνδυασμό φύλου-χρώματος αν η υπόθεση του δημοσκόπου ήταν αληθινή.

(γ) Να βρείτε αν η κατανομή των αποκρίσεων είναι συμβατή με τον ισχυρισμό του δημοσκόπου ($α = 0,05$).

Σημείωση: Δίνεται ότι αν $\chi^2 \sim \chi^2(2)$, τότε $P(\chi^2 > 5) = 0,082$.

Βαθμολογία: (α) 0,4μ β) 0,8μ (γ) 0,8μ

Θέμα 4. Ένας δάσκαλος ισχυρίζεται ότι υπάρχει σημαντική διαφοροποίηση μεταξύ αγοριών και κοριτσιών ως προς την επίδοσή τους στη γλώσσα. Εξετάζει τους μαθητές του (64 αγόρια και 25 κορίτσια) σε μία τυποποιημένη γραπτή δοκιμασία και βγάζει τα εξής αποτελέσματα:

Αγόρια: $n_1 = 64$, $x_1 = 74{,}8$, $s_1 = 3{,}8$. Κορίτσια: $n_2 = 25$, $x_2 = 81{,}4$, $s_2 = 3{,}5$.

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 μέση επίδοση των αγοριών.

Βαθμολογία: (α) 0,4μ (β) 1,9μ (γ) 0,7μ

Τυπολόγιο

Α. Πίνακας κανονικής κατανομής: $\Phi(z) = P(Z < z)$, για $2 \le z \le 2{,}99$ όπου $Z \sim N(0, 1)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861

Σημείωση: Για τιμές του z ανάμεσα σε δύο τιμές του πίνακα υπολογίστε το ημιάθροισμα των αντίστοιχων πιθανοτήτων.

Β. Κανόνας του Bayes: $P(A \mid B) = P(B \mid A)P(A) / P(B)$

Γ. Κατανομές
$X \sim N(\mu, \sigma^2) \rightarrow EX = \mu$, $VarX = \sigma^2$. $X \sim t(n)$ και $n > 30 \rightarrow X \sim N(0, 1)$.
$X \sim Poisson(\lambda)$, $P(X = k) = e^{-\lambda} \lambda^k/k!$, $EX = VarX = \lambda$ και για λ αρκετά μεγάλο $X \sim N(\lambda, \lambda)$.
$X \sim B(n, p)$, $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$, $EX = np$, $VarX = npq$ και $X \sim N(np, npq)$ ($np > 5, npq > 5$).
Δοκιμασία χ^2 ως έλεγχος ανεξαρτησίας: $\chi_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(O_{ii} - E_{ii})^2}{E_{ii}} \sim \chi^2((k-1)\cdot(m-1))$

Δ. Για τη διαφορά των δειγματικών μέσων δύο ανεξάρτητων δειγμάτων γνωρίζουμε ότι το τυπικό σφάλμα υπολογίζεται ως $s_p = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

(β) Αν $\overline{X}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$ και $\overline{X}_2 \sim N(\mu_2, \frac{\sigma^2}{n_2})$ τότε $t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{s_p} \sim t(n_1 + n_2 - 2)$

(γ) 95% διάστημα εμπιστοσύνης $\left(\overline{X} - \frac{s}{\sqrt{n}}1{,}96, \overline{X} + \frac{s}{\sqrt{n}}1{,}96\right)$, $n \ge 30$.

Ενδεικτικές απαντήσεις

Θέμα 1

Ορίζουμε A = {το μήνυμα είναι Spam}, B = {το μήνυμα περιέχει τη λέξη "Rolex"}.

Από τα 3000 μηνύματα, τα 2000 είναι Spam και τα 1000 δεν είναι Spam, άρα

$$P(A) = 2.000/3.000 = 2/3.$$

Η λέξη «Rolex» εμφανίστηκε σε 250 από τα 2000 μηνύματα που είναι Spam, άρα

$$P(B \mid A) = 250/2.000 = 1/8.$$

Η λέξη «Rolex» εμφανίστηκε σε 5 από τα 1000 μηνύματα που δεν είναι Spam, άρα

$$P(B \mid A') = 5/1.000 = 1/200.$$

Από το νόμο ολικής πιθανότητας υπολογίζουμε

$P(B) = P(B \mid A)P(A) + P(B \mid A')P(A') = 1/8 \cdot 2/3 + 1/200 \cdot 1/3 = 1/12 + 1/600 = 51/600.$

α) P({το μήνυμα αυτό είναι spam δεδομένου ότι περιέχει τη λέξη «Rolex»})
$= P(A \mid B) = P(B \mid A)P(A) / P(B) = 1/8 \cdot 2/3 / (51/600) = 2 \cdot 600 / (24 \cdot 50) = 0{,}98 = 98\%$

β) P({το μήνυμα περιέχει τη λέξη "Rolex" δεδομένου ότι δεν είναι spam})
$= P(B \mid A') = 1/200.$

γ) P({το μήνυμα περιέχει τη λέξη "Rolex" και είναι spam}) = $P(A \cap B)$
$= P(B \mid A)\, P(A) = 1/8 \cdot 2/3 = 1/12.$

Θέμα 2

Έστω X = {πλήθος εξαρτημάτων που κατασκευάζονται σε μία ημέρα}. Κάθε μία κατασκευή κατασκευάζεται ανεξάρτητα από τις άλλες, άρα X ~ Poisson(7) και $P(X = k) = e^{-7}\, 7^k/k!$.

α) $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = e^{-7}\,(1 + 7 + 49/2) = 0{,}0296 = 2{,}96\%$.

Έστω Y = {πλήθος εξαρτημάτων μεταξύ των 9 που αντέχουν στην καταπόνηση}. Είναι Y ~ B(9, 0.8) και $P(X = k) = (9 \text{ ανά } k) 0{,}8^k \cdot 0{,}2^{9-k}$.

β) P({αντέχουν το πολύ 7 εξαρτήματα}) = 1 - P({αντέχουν 8 ή 9 εξαρτήματα})
$= 1 - P(Y = 8) - P(Y = 9)$
$= 1 - (9 \text{ ανά } 8) 0{,}8^8 \cdot 0{,}2 + (9 \text{ ανά } 9) 0{,}8^9$
$= 0{,}436 = 43{,}6\%$.

γ) P({αντέχουν λιγότερα από 6 και τουλάχιστον 4 εξαρτήματα})
= P({αντέχουν 4 ή 5 εξαρτήματα})
$= P(Y = 4) + P(Y = 5)$
$= (9 \text{ ανά } 4) 0{,}8^4 \cdot 0{,}2^5 + (9 \text{ ανά } 5) 0{,}8^5 \cdot 0{,}2^4$
$= 0{,}0826 = 8{,}26\%$.

Θέμα 3

α) Είναι η δοκιμασία χ^2 ως έλεγχος ανεξαρτησίας

H_0: το φύλο είναι ανεξάρτητο από το επιλεγμένο χρώμα και H_1: όχι η H_0.

β) Αναμενόμενες συχνότητες υπό την υπόθεση H_0

Χρώμα	Κόκκινο	Κίτρινο	Μπλε
Άνδρες	29	32,3	38,7
Γυναίκες	58	64,7	77,3

γ) $\chi_0^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ii}-E_{ii})^2}{E_{ii}} = 5$, και $\chi_0^2 \sim \chi^2((3–1)\cdot(2-1)) = \chi^2(2)$.

Είναι $p = P(\chi^2 > \chi_0^2) = P(\chi^2 > 5) = 0{,}082 > 0{,}05$, άρα δεν απορρίπτεται η υπόθεση H_0, δηλαδή η κατανομή των αποκρίσεων δεν είναι συμβατή με τον ισχυρισμό του δημοσκόπου.

Θέμα 4

α) Είναι η δοκιμασία t – test για δύο ανεξάρτητα δείγματα.

H_0: $\mu_A = \mu_K$ και H_1: $\mu_A \neq \mu_K$

β) Υπολογίζουμε $s_p = \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = 0{,}877$ και

$t_0 = \frac{\overline{X}_1-\overline{X}_2}{s_p} = 7{,}523$.

Είναι $t \sim t(64 + 25 – 2) = t(87) \sim N(0, 1)$ και

$p = P(t > t_0) = P(t > 7{,}523) = 1 – P(t \leq 7{,}523) = 1 – 1 = 0 < \alpha = 0{,}05$, άρα η μηδενική υπόθεση απορρίπτεται.

Συμπεραίνουμε ότι οι δύο ομάδες έχουν σημαντικά διαφορετική μέση επίδοση στη Γλώσσα.

γ) 95% δ.ε. για τη μέση επίδοση των αγοριών

$$\left(\overline{X}-\frac{s}{\sqrt{n}}1{,}96,\ \overline{X}+\frac{s}{\sqrt{n}}1{,}96\right) = (73{,}9,\ 75{,}7).$$