Μάθημα: Στατιστική (Κωδ. 105)

Διδάσκων: Γιώργος Κ. Παπαδόπουλος

8. Έλεγχοι X^2

Σύντομη ανασκόπηση βασικών εννοιών, προτάσεων και τύπων

Έλεγχος X^2 καλής προσαρμογής

Έστω ότι μια πειραματική/δειγματοληπτική μονάδα μπορεί να ταξινομηθεί ως προς **ένα** χαρακτηριστικό σε ακριβώς μια από k κατηγορίες $r_1, r_2, \ldots, r_k$ και έστω p_i η πιθανότητα ταξινόμησής της στην κατηγορία r_i, $i = 1, 2, \ldots, k$. Αν ν πειραματικές/δειγματοληπτικές μονάδες ταξινομήθηκαν ως προς το χαρακτηριστικό αυτό και οι O_1 ταξινομήθηκαν στην κατηγορία r_1, οι O_2 στην κατηγορία r_2, … και οι O_k στην κατηγορία r_k, τότε σε επίπεδο σημαντικότητας α, η μηδενική υπόθεση

$$H_0 : p_1 = p_{10},\ p_2 = p_{20}, \ldots, p_k = p_{k0}$$

όπου $p_{10}, p_{20}, \ldots, p_{k0}$ **γνωστές** πιθανότητες με $\sum_{i=1}^{k} p_{i0} = 1$

απορρίπτεται έναντι της εναλλακτικής

$$H_1 : p_i \neq p_{i0} \text{ για ένα τουλάχιστον } i,\ i = 1, 2, \ldots, k$$

αν

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \geq \chi^2_{k-1;\alpha}$$

και εφόσον $E_i = \nu p_{i0} \geq 5$, για κάθε $i = 1, 2, \ldots, k$. Με $E_1, E_2, \ldots, E_k$ συμβολίζουμε τις ***αναμενόμενες συχνότητες*** των αποτελεσμάτων $r_1, r_2, \ldots, r_k$ αν θεωρήσουμε ότι η μηδενική υπόθεση είναι αληθής.

Αν οι πιθανότητες $p_{10}, p_{20}, \ldots, p_{k0}$ **δεν είναι γνωστές**, εκτιμώνται από τις $\hat{p}_{i0}$ (με βάση το δείγμα) και η απορριπτική περιοχή του ελέγχου σε επίπεδο σημαντικότητας α ορίζεται από την ανισότητα

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \geq \chi^2_{k-1-m;\alpha}$$

και εφόσον $\hat{E}_i = \nu \hat{p}_{i0} \geq 5$, για κάθε $i = 1, 2, \ldots, k$. Με m συμβολίζουμε τον αριθμό των παραμέτρων που πρέπει να εκτιμηθούν για να μπορούν να εκτιμηθούν οι αναμενόμενες συχνότητες.

Έλεγχος X^2 ανεξαρτησίας

Έστω ότι μια πειραματική/δειγματοληπτική μονάδα μπορεί να ταξινομηθεί ως προς **δύο** χαρακτηριστικά A και B από τα οποία το A μπορεί να πάρει $r \geq 2$ διαφορετικές τιμές (κατηγορίες) $A_1, A_2, \ldots, A_r$ και το B μπορεί να πάρει $c \geq 2$ διαφορετικές τιμές (κατηγορίες) $B_1, B_2, \ldots, B_c$. Έστω δηλαδή ότι μια πειραματική/δειγματοληπτική μονάδα μπορεί να ταξινομηθεί σε ακριβώς μια από rc διαφορετικές κατηγορίες (A_i, B_j). Αν ν πειραματικές/δειγματοληπτικές μονάδες ταξινομήθηκαν ως προς τα δύο αυτά χαρακτηριστικά A και B και O_{ij} από αυτές ταξινομήθηκαν στην κατηγορία (A_i, B_j), τότε σε επίπεδο σημαντικότητας α, η μηδενική υπόθεση

$$H_0 : \textit{ τα χαρακτηριστικά A και B είναι ανεξάρτητα}$$

απορρίπτεται έναντι της εναλλακτικής

$$H_1 : \textit{ τα χαρακτηριστικά A και B δεν είναι ανεξάρτητα}$$

αν

$$X^2 = \sum_{\forall i,j} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \geq \chi^2_{(r-1)(c-1);\alpha}$$

και εφόσον $\hat{E}_{ij} \geq 5$ για όλα τα i και j. Με $\hat{E}_{ij}$ συμβολίζουμε την ***εκτιμώμενη αναμενόμενη συχνότητα*** της κατηγορίας (A_i, B_j) αν θεωρήσουμε ότι η μηδενική υπόθεση είναι αληθής και είναι $\hat{E}_{ij} = \frac{R_i C_j}{\nu}$ όπου, R_i η παρατηρηθείσα συχνότητα της κατηγορίας A_i και C_j η παρατηρηθείσα συχνότητα της κατηγορίας B_j.

Προβλήματα και Ασκήσεις

1. Στη βιβλιογραφία αναφέρεται ότι τα ποσοστά των ομάδων αίματος Α, Β, ΑΒ και Ο σε έναν πληθυσμό είναι 0.41, 0.10, 0.04 και 0.45 αντίστοιχα. Μια ομάδα ερευνητών, προκειμένου να ελέγξει αν τα ποσοστά των ομάδων αίματος σε αυτόν τον πληθυσμό είναι πράγματι αυτά που αναφέρονται στην βιβλιογραφία, επέλεξε τυχαία 200 άτομα από αυτόν τον πληθυσμό και για καθένα κατέγραψε την ομάδα αίματός του. Στον πίνακα που ακολουθεί φαίνεται η συχνότητα κάθε ομάδας αίματος που παρατηρήθηκε στο δείγμα.

Ομάδα αίματος

	A	B	AB	O
Παρατηρηθείσα συχνότητα	89	18	12	81

Σε επίπεδο σημαντικότητας 5%, τα ποσοστά που παρατηρούνται στο δείγμα συμφωνούν ή όχι, με τα αντίστοιχα ποσοστά που αναφέρονται στη βιβλιογραφία;

2. Ένας ερευνητής σχεδίασε και εκτέλεσε το εξής πείραμα. Σε έναν κλειστό διάδρομο στο τέλος του οποίου υπήρχαν τρεις έξοδοι διαφορετικού χρώματος (πράσινη, κόκκινη και μπλε), απελευθέρωσε ένα ποντίκι 90 φορές και κατέγραψε πόσες φορές αυτό διέφυγε από την πράσινη έξοδο, πόσες από την κόκκινη και πόσες από τη μπλε. Η συχνότητα που παρατηρήθηκε για κάθε έξοδο διαφυγής φαίνεται στον πίνακα που ακολουθεί.

Έξοδος διαφυγής

	Πράσινη	Κόκκινη	Μπλε
Παρατηρηθείσα συχνότητα	20	39	31

Σε επίπεδο σημαντικότητας 5%, υποστηρίζουν αυτά τα πειραματικά δεδομένα ότι το ποντίκι δε δείχνει την ίδια προτίμηση και για τις τρεις εξόδους;

3. Από κατάλληλη διασταύρωση φυτών πετούνιας προκύπτουν, ως προς το χρώμα του άνθους και το σχήμα του φύλλου, οι εξής τέσσερις τύποι φυτών: AB (κόκκινα άνθη και στρογγυλά φύλλα), Ab (κόκκινα άνθη και μακρόστενα φύλλα), aB (λευκά άνθη και στρογγυλά φύλλα) και ab (λευκά άνθη και μακρόστενα φύλλα). Σύμφωνα με το μοντέλο κληρονομικότητας του Mendel, οι τέσσερις τύποι απογόνων, AB, Ab, aB και ab πρέπει να βρίσκονται σε αναλογία 9:3:3:1. Σε ένα σχετικό πείραμα, από 160 πειραματικά φυτά τα 95 βρέθηκαν να είναι τύπου AB, τα 30 τύπου Ab, τα 28 τύπου aB και τα 7 τύπου ab. Σε επίπεδο σημαντικότητας 1%, αυτά τα πειραματικά δεδομένα δίνουν άραγε σημαντικές αποδείξεις εναντίον του μοντέλου κληρονομικότητας του Mendel;

4. Ρωτήσαμε καθέναν από 200 ενήλικες κατοίκους μιας μικρής επαρχιακής πόλης τους οποίους επιλέξαμε με βάση ένα σχέδιο τυχαίας δειγματοληψίας, αν είναι καπνιστής, περιστασιακός καπνιστής, πρώην καπνιστής ή μη καπνιστής.

Συνήθεια ως προς το κάπνισμα

		Καπνιστής	Περιστασιακός καπνιστής	Πρώην καπνιστής	Μη καπνιστής
Φύλο	Άνδρας	40	14	10	60
	Γυναίκα	20	4	4	48

Με βάση αυτά τα δεδομένα, να ελέγξετε σε επίπεδο σημαντικότητας 5%, αν υπάρχει συνάφεια/εξάρτηση μεταξύ φύλου και συνήθειας των κατοίκων της συγκεκριμένης πόλης ως προς το κάπνισμα.

5. Στο πλαίσιο μιας έρευνας αγοράς, ρωτήθηκε καθένας από 670 δυνητικούς αγοραστές ενός νέου προϊόντος να δηλώσει ποιο από τρία μοντέλα (Μ1, Μ2, Μ3) του νέου προϊόντος προτιμά καθώς και να κατατάξει τον εαυτό του σε έναν από τέσσερις τύπους καταναλωτή που του προτάθηκαν (Τ1, Τ2, Τ3, Τ4).

Τύπος καταναλωτή

		T1	T2	T3	T4
Μοντέλο	M1	20	22	56	38
	M2	40	44	68	42
	M3	80	90	75	95

Με βάση αυτά τα δεδομένα να ελέγξετε σε επίπεδο σημαντικότητας 5%, αν υπάρχει εξάρτηση μεταξύ τύπου καταναλωτή και προτίμησης μοντέλου.

6. Μια ομάδα ερευνητών για να μελετήσει την εξάπλωση μιας ασθένειας *(υφέρπουσα σήψη)* στις φυτικές καλλιέργειες, διαίρεσε μια καλλιέργεια λάχανου σε 270 τετράγωνα τμήματα καθένα από τα οποία περιείχε τον ίδιο αριθμό λάχανων και κατέγραψε, ανά τμήμα, τον αριθμό των φυτών που παρουσίαζαν σημάδια *υφέρπουσας σήψης*. Τα αποτελέσματα των μετρήσεων στα 270 τμήματα της καλλιέργειας δίνονται στον πίνακα συχνοτήτων που ακολουθεί.

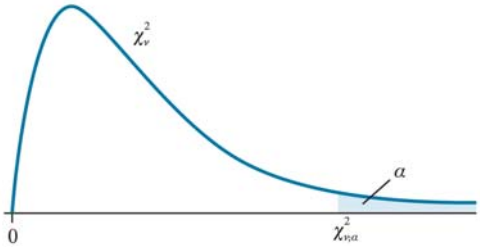
Αριθμός φυτών που έχουν προσβληθεί/τμήμα	**Παρατηρηθείσα συχνότητα**
0	38
1	57
2	68
3	47
4	23
5	9
6	10
7	7
8	3
9	4
10	2
≥ 11	2
Σύνολο	270

Να ελέγξετε, σε επίπεδο σημαντικότητας 5%, αν ο αριθμός των φυτών που παρουσιάζουν σημάδια *υφέρπουσας σήψης* (ανά τμήμα) περιγράφεται από μια κατανομή *Poisson*. Πώς μπορούμε να εξηγήσουμε το αποτέλεσμα αυτού του ελέγχου;

7. Από το αρχείο ενός μεγάλου μαιευτηρίου επιλέξαμε, με βάση ένα σχέδιο τυχαίας δειγματοληψίας, 70 ιατρικούς φακέλους εγκύων και από κάθε φάκελο καταγράψαμε τη διάρκεια κύησης (σε ημέρες). Τα δεδομένα που προέκυψαν φαίνονται στον πίνακα που ακολουθεί.

251	264	234	283	226	244	269	241	276	274
263	243	254	276	241	232	260	248	284	253
265	235	259	279	256	256	254	256	250	269
240	261	263	262	259	230	268	284	259	261
268	268	264	271	263	259	294	259	263	278
267	293	247	244	250	266	286	263	274	253
281	286	266	249	255	233	245	266	265	264

α) Υποστηρίζουν αυτά τα δειγματοληπτικά δεδομένα ότι η διάρκεια της κύησης ακολουθεί μια κανονική κατανομή; Να κάνετε κατάλληλο στατιστικό έλεγχο σε επίπεδο σημαντικότητας 0.05 β) Σε μια πρόσφατη δημοσίευση αναφέρεται ότι ο μέσος και η τυπική απόκλιση της διάρκειας της κύησης είναι 266 και 16 ημέρες αντίστοιχα. Τα συγκεκριμένα δειγματοληπτικά δεδομένα υποστηρίζουν άραγε ότι πράγματι έτσι είναι, δηλαδή, ότι η διάρκεια της κύησης ακολουθεί κανονική κατανομή με $\mu = 266$ και $\sigma = 16$; Να κάνετε κατάλληλο στατιστικό έλεγχο σε επίπεδο σημαντικότητας 0.05.



Ο πίνακας δίνει τα σημεία $\chi^2_{\nu;\alpha}$ για τα οποία

$$P(X > \chi^2_{\nu;\alpha}) = \alpha$$

με $X \sim \chi^2_\nu$.

ν	$\alpha = 0.995$	$\alpha = 0.99$	$\alpha = 0.975$	$\alpha = 0.95$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.00$
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.414	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.706	22.164	24.433	26.509	55.756	59.342	63.691	66.766
50	27.991	29.708	32.357	34.764	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.392	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.930	124.342	129.561	135.807	140.169