ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2020-2021

3ο ΣΕΤ ΑΣΚΗΣΕΩΝ

ΓΡΗΓΟΡΙΑΔΟΥ ΓΕΩΡΓΙΑ ,3160029

ΑΣΚΗΣΗ 1 (α)

Πραγματοποιούμε άπειρες ρίψεις και παίρνουμε δείγμα από αυτές με n = 50 ρίψεις. Έστω X οι κορώνες και Y τα γράμματα με X = 29 και Y = 21.

X > 15 και Y >15 άρα τα δεδομένα μας είναι κατάλληλα.

$\hat{p} = \frac{X}{n} = \frac{29}{50} = 0.58$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.58*0.42}{50}} = 0.069$

Για διάστημα εμπιστοσύνης C=95% έχουμε z_*=1.96 οπότε το ζητούμενο διάστημα είναι $\hat{p} \pm z_* \sigma_{\hat{p}} =$ [0.4431 , 0.7168]

ΑΣΚΗΣΗ 1 (β)

Δίπλευρος έλεγχος H_0 : p = 0.5 , Αφού το ζάρι είναι δίκαιο άρα η συχνότητα εμφάνισης κορώνας είναι ½.

Στατικός έλεγχος z = $\frac{0.58-0.5}{\sqrt{\frac{0.5(1-0.5)}{50}}}$ = 1.1313 άρα p-value = 2Φ(-|z|) = 0.25. Το p-value είναι μικρή τιμή άρα δε θα μπορούσαμε να απορρίψουμε την μηδενική υπόθεση με σιγουριά.

ΑΣΚΗΣΗ 1 (γ)

N ≥ $\frac{z_*^2}{4m^2}$ = $\frac{1.96^2}{4*0.01^2}$ = 9604   Άρα θα έπρεπε να πραγματοποιήσει 9604 ρίψεις.

ΑΣΚΗΣΗ 2

Το μέγεθος του δείγματος που χρησιμοποιείται για τις δημοσκοπήσεις είναι ανεξάρτητο από το μέγεθος του πληθυσμού, καθώς σύμφωνα με τον τύπο $n \geq \frac{z_*^2}{4m^2}$ το πρώτο εξαρτάται μόνο από το περιθώριο λάθους m και το επίπεδο εμπιστοσύνης. Άρα και για τις δημοσκοπήσεις στις Η.Π.Α ο αριθμός των 1100 ατόμων είναι αρκετός.

ΑΣΚΗΣΗ 3 (α)

Έστω p_1 το ποσοστό των καπνιστών που είναι άνδρες και p_2 που είναι γυναίκες.

Δίπλευρος έλεγχος $H_0 : p_1 = p_2$

Έχουμε $n_1 = 30$ το πλήθος των αντρών και $n_2 = 30$ το πλήθος των γυναικών.

Επίσης $\widehat{p_1}$= 12/30 = 0.4 το ποσοστό των αντρών που καπνίζουν και $\widehat{p_2}$= 14/30 = 0.46 το ποσοστό των γυναικών που καπνίζουν. Επίσης $\hat{p} = \frac{12+14}{60}$ = 0.43

Στατιστικός έλεγχος z = $\frac{0.4-0.46}{\sqrt{\frac{0.43(1-0.4)}{30}+\frac{0.43(1-0.46)}{30}}}$ = -0.521

Άρα το p-value = 0.60 το οποίο είναι αρκετά μεγάλο οπότε δε μπορούμε να απορρίψουμε τη μηδενική υπόθεση . Αυτό σημαίνει ότι υπάρχει σχέση μεταξύ φύλου και καπνίσματος.

ΑΣΚΗΣΗ 3 (β)

Το ζητούμενο διάστημα εμπιστοσύνης είναι :

$$\widehat{p_1} - \widehat{p_2} \pm z_* \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n2}} = 0.4 - 0.46 \pm 1.96\sqrt{\frac{0.4(1-0.4)}{30} + \frac{0.46(1-0.46)}{30}} =$$

[−0.194,0.061]

ΑΣΚΗΣΗ 3 (γ)

Έστω ο έλεγχος H_0 : Το φύλο δεν έχει σχέση με το κάπνισμα

H_a : Το φύλο έχει σχέση με το κάπνισμα

ΠΙΝΑΚΑΣ ΣΥΝΑΦΕΙΑΣ

ΦΥΛΟ	**ΚΑΠΝΙΣΤΗΣ**	**ΜΗ ΚΑΠΝΙΣΤΗΣ**	
ΑΝΤΡΕΣ	12	18	30
ΓΥΝΑΙΚΕΣ	14	16	30
	26	34	60

ΑΣΚΗΣΗ 3 (δ)

ΦΥΛΟ	**ΚΑΠΝΙΣΤΗΣ**	**ΜΗ ΚΑΠΝΙΣΤΗΣ**	
ΑΝΤΡΕΣ	13	17	30
ΓΥΝΑΙΚΕΣ	13	17	30
	26	34	60

Στατιστικός έλεγχος $X^2 = \frac{(12-13)^2}{13} + \frac{(18-17)^2}{17} + \frac{(14-13)^2}{13} + \frac{(16-17)^2}{17} \approx 0.07$ $+0.058 + 0.07 + 0.058 \approx 0.256$

ΑΣΚΗΣΗ 4 (α)

Θέλουμε να μάθουμε αν παρασκευάζονται περισσότερα κόκκινα smarties από ότι μπλε, άρα μας ενδιαφέρει ο υποπληθυσμός των κόκκινων και μπλε smarties. Έστω p το ποσοστό των κόκκινων και 1-p των μπλε κουφέτων.

Ενδιαφερόμαστε να μάθουμε αν τα κόκκινα smarties είναι περισσότερα οπότε θα πάρουμε μονόπλευρο έλεγχο υπόθεσης H_0 : p = $\frac{1}{2}$, H_a : p > $\frac{1}{2}$.

Από τα δεδομένα της εκφώνησης έχουμε ότι:
n = 19 + 15 = 34, X = 19 και $\hat{p} = \frac{X}{n} = \frac{19}{34}$ = 0.5588.

Θεωρούμε ότι η συσκευασία αποτελεί ένα τυχαίο δείγμα με τα smarties που φτιάχνονται, δηλαδή ότι ο τρόπος με τον οποίο αναμιγνύονται τα χρώματα είναι τυχαίος.

Ο στατιστικός έλεγχος είναι $z = \frac{\hat{p}-0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}}$ = 0.686 και το p-value = 2Φ(-|z|)=0.4902.

Αφού η τιμή του p-value δεν είναι πολύ μικρή, η μηδενική υπόθεση είναι αποδεκτή.
Άρα, δεν υπάρχει σημαντική διαφορά στον αριθμό των κόκκινων και μπλε κουφέτων.

ΑΣΚΗΣΗ 4 (β)

Θέλουμε να δούμε αν η κατανομή των χρωμάτων είναι η ίδια σε σχέση με αυτή του 2009.
Θα εφαρμόσουμε τον χ^2 έλεγχο καλής προσαρμογής:

H_0: η κατανομή των χρωμάτων καφέ, κόκκινο, κίτρινο, μπλε και πράσινο είναι 19.8%, 17.8%, 17.6%, 19.6% και 25.2% αντίστοιχα,
H_α: η κατανομή των χρωμάτων είναι διαφορετική.

Από τα δεδομένα προκύπτει ο ακόλουθος πίνακας:

	Δεδομένα	Αναμενόμενες Τιμές
Καφέ	22	15.84
Κόκκινο	19	14.24
Κίτρινο	16	14.08
Μπλέ	15	15.68
Πράσινο	8	20.16

Με τη βοήθεια της R βρίσκουμε ότι p-value= 0.02, το οποίο είναι αρκετά μικρό για να ισχύει η μηδενική υπόθεση. Άρα η κατανομή των χρωμάτων είναι διαφορετική σε σχέση με το 2009.

<u>ΑΣΚΗΣΗ 4 (γ)</u>

Θα θεωρήσουμε ότι η συσκευασία των M&Ms αποτελεί ένα τυχαίο δείγμα από τον πληθυσμό των κουφέτων της συγκεκριμένης μάρκας. Θα ελέγξουμε αν τα δείγματα των smarties και των M&Ms προήλθαν από πληθυσμούς με την ίδια κατανομή χρωμάτων. Για αυτό τον έλεγχο θα χρησιμοποιήσουμε έναν χ^2 έλεγχο για ομοιογένεια.

H_0: οι πληθυσμοί των smarties και των M&Ms είναι ομοιογενείς.
H_α: οι πληθυσμοί δεν είναι ομοιογενείς.

Από τα δεδομένα προκύπτει ο ακόλουθος πίνακας συνάφειας:

	smarties	M&Ms	
Καφέ	22	10	32
Κόκκινο	19	12	31
Κίτρινο	16	20	36
Μπλέ	15	9	24
Πράσινο	8	5	13
	80	56	136

Με τη βοήθεια της R βρίσκουμε ότι ο στατιστικός έλεγχος είναι χ^2 = 4.626, οι βαθμοί ελευθερίας = 4 και p-value= 0.3278.

Το p-value είναι αρκετά μεγάλο ώστε να μην απορριφθεί η μηδενική υπόθεση, οπότε οι κατανομές των χρωμάτων στα smarties και τα M&Ms είναι ίδιες.