3η Σειρά Ασκήσεων Στατιστικής

Ειρήνη Χρυσικοπούλου(3180208)
Παναγιώτης Παναγιώτου(3180139)

January 2022

Άσκηση 1

a.

Αρχικά το δείγμα είναι αρκετά μεγάλο,$n = 50$, και έχει ληφθεί με τυχαίο τρόπο.Επίσης $X_1 = 29 \geq 15$ και $50 - X_1 = 21 \geq 15$ επομένως το διάστημα εμπιστοσύνης που θα βρούμε θα έχει καλή ακρίβεια.
$z_* = 1.959964$
δειγματικό ποσοστό εμφάνισης κορώνας$= \hat{p} = \frac{29}{50} = 0.58$
ο τύπος για το διάστημα επιστοσύνης είναι: $\hat{p} \pm z_* \sqrt{(\hat{p} - (1 - \hat{p})/n} \iff$
$0.58 \pm 1.959964 * 0.07 = [0.4428025, 0.7171975]$

b.

Εκτελούμε τον δίπλευρο έλεγχο $H_0 : p = 0, 5$,όπου p οι εμφανίσεις της κορώνας αν κάνουμε άπειρες ρίψεις του νομίσματος. Το στατιστικό z ισούται με
$z = \frac{0.58-0.5}{\sqrt{\frac{0.5^2}{n}}} = 1.131371$
Επομένως το $pvalue$ ισούται με $pvalue = 0.257899$
Συνεπώς αφού $pvalue > \alpha = 0.05$ δεν απορρίπτουμε την μηδενική. Επομένως το νόμισμα είναι δίκαιο.

c.

Θα θέλαμε $z_* \sqrt{\frac{(\hat{p}-(1-\hat{p})}{n}} \leq 0.01 \iff$
$n \geq \frac{z_*^2 \hat{p}(1-\hat{p})}{0.01^2} = 9357.794$
Επομένως θα χρειαζόμασταν 9358 ρίψεις.

Άσκηση 2η

Όταν δεν γνωρίζουμε το το ποσοστό εμφάνισης p μιας τιμής μιας κατηγορικής μεταβλητής στον πληθυσμό,ούτε το δειγματικό ποσοστό $\hat{p}$ μπορούμε να χρησιμοποιήσουμε την σχέση $p(1 - p) \leq 1/4 \Rightarrow n \geq \frac{z_*^2}{4m^2}$,όπου $m =$ περιθώριο σφάλματος $= 0.03$
$n \geq \frac{z_*^2}{4m^2} \iff n = 1067.072$
Επομένως θα μπορούσαμε να κατασκευάσουμε διάστημα επιστοσύνης 95% με περιθώριο σφάλματος 3% με δείγμα μεγέθους $n = 1067.072$.
Παρατηρούμε ότι το μέγεθος του πληθυσμού δεν επηρεάζει το μέγεθος του δείγματος επομένως για τόσο για τις δημοσκοπήσεις στην Ελλάδα όσο και για τις δημοσκοπήσεις στις ΗΠΑ μπορεί να χρησιμοποιηθεί το ίδιο μέγεθος δείγματος.

Άσκηση 3η

a.

Επειδή $X_1 = 14 \geq 5, n_1 - X_1 = 16 \geq 5, X_2 = 12, n_2 - X_2 = 18$ και έχουμε μεγάλο δείγμα ,$n_1 = 30, n_2 = 30$, για τον υποπληθυσμό των ανδρών και των γυναικών αντίστοιχα το $pvalue$ είναι ακριβές.
Για τον πληθυσμό των γυναικών έχουμε:
Μέγεθος δείγματος: $n_1 = 30$
Πλήθος γυναικών που καπνίζουν: $X_1 = 14$
Δειγματικό Ποσοστό γυναικών που καπνίζουν:$\hat{p}_1 = \frac{14}{30} = 0.467$

Για τον πληθυσμό των ανδρών έχουμε:
Μέγεθος δείγματος: $n_2 = 30$
Πλήθος ανδρών που καπνίζουν: $X_2 = 12$
Δειγματικό Ποσοστό ανδρών που καπνίζουν: $\hat{p}_2 \frac{12}{30} = 0.40$

Θα εκτελέσουμε τον δίπλευρο έλεγχο σημαντικότητας της μηδενικής υπόθεσης $H_0 : p_1 = p_2$ όπου p_1, p_2 τα ποσοστά καπνιστών στον πληθυσμό των γυναικών και των ανδρών αντίστοιχα. Στατιστικό ελέγχου:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}, \text{ όπου } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Επομένως $z = 0.5210501$
Υπολογισμός του $pvalue$ στην R:
2*pnorm(-abs(z))
0.6023319 // $pvalue$

Παρατηρούμε ότι το $pvalue$ είναι μεγάλο.Επομένως δεν μπορούμε αν απορρίψουμε την μηδενική υπόθεση,ότι το ποσοστό καπνιστριών είναι ίσο με αυτό των καπνιστών.Επομένως δεν υπάρχει σχέση ανάμεσα στο κάπνισμα και το φύλο.

b.

Ο τύπος για ένα $C\%$ διάστημα επιστοσύνης είναι:

$$C\%: \hat{p}_1 - \hat{p}_2 \pm z_* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$z_* = 1.959964$,επομένως το διάστημα επιστοσύνης είναι το:
$[-0.1835364, 0.3168697]$

c.

Πίνακας συνάφειας:

		Φύλο		
		Γυναίκα	Άντρας	
Καπνίζει	TRUE	14	12	26
	FALSE	16	18	34
		30	30	60

χ^2 έλεγχος σημαντικότητας:
H_0: το φύλο και το κάπνισμα είναι ανεξάρτητα
H_α: το φύλο και το κάπνισμα δεν είναι ανεξάρτητα

d.

Εαν ισχύει η μηδενική υπόθεση τότε ο πίνακας αναμενώμεων τιμών είναι:

		Φύλο		
		Γυναίκα	Άντρας	
Καπνίζει	TRUE	(26*30)/60=13	(26*30)/60=13	26
	FALSE	(34*30)/60=17	(34*30)/60=17	34
		30	30	60

Επομένως το στατιστικό χ^2 είναι:$\chi^2 = \frac{(14-13)^2}{13} + \frac{(16-17)^2}{17} + \frac{(12-13)^2}{13} + \frac{(18-17)^2}{17} =$ 0.2714932

Εύρεση στατιστικού χ^2 και $pvalue$ με εντολή της R:
chisq.test(tps,correct=FALSE)

Pearson's Chi-squared test

data: tps
X-squared = 0.27149, df = 1, p-value = 0.6023

Επομένως $pvalue = 0.6023$,άρα δεν υπάρχει συσχέτιση ανάμεσα στο φύλο και το κάπνισμα. Παρατηρούμε ότι το $pvalue$ που βρήκαμε το ερώτημα (α) είναι το ίδιο με αυτό που βρήκαμε σε αυτό το ερώτημα.Επομένως οι δύο έλεγχοι (δηλαδή ο έλεγχος χ^2 και ο έλεγχος z) είναι ισοδύναμοι(αυτό συμβαίνει πάντα στους 2x2 πίνακες).

Άσκηση 4η

a.

Το μέγεθος του δείγματος ισούται με $n = 34$,γιατί μελατάμε τον υποπληθυσμό των μπλε και κόκκινων smarties.
Για τα κόκκινα smarties έχουμε:
Αριθμός κόκκινων smarties στο δείγμα: $X_1 = 19$
Δειγματικό ποσοστό : $p_1 = \hat{1}9/34$

Για τα μπλε smarties έχουμε:
Αριθμός μπλε smarties στο δείγμα: $X_2 = 15$
Δειγματικό ποσοστό : $\hat{p_2} = 15/34$

Θέλουμε ελέξουμε αν σε ένα πληθυσμό που αποτελείται μόνο από κόκκινα και μπλε smarties το ποσοστό εμφάνισης των κόκκινων είναι ίδιο με το ποσοστό εμφάνισης των μπλε.Έστω p το ποσοστό εμφάνισης των κόκκινων smarties.Επειδή ο πληθυσμός που μελετάμε περιέχει μόνο μπλε και κόκκινα smarties το ποσοστό εμφάνισης των μπλε είναι $1-p$. Παρασκευάζεται ίδια ποσότητα μπλε και κόκκινων smarties αν και μόνο αν $p = \frac{1}{2}$.
Θα εφαρμόσουμε τον έλεγχο σημαντικότητας $H_0 : p = \frac{1}{2}$,
$H_\alpha : p > \frac{1}{2}$(αυτό γιατί μας ενδιαφέρει μόνο αν οι κόκκινες smarties είναι περισσότερες από τις μπλε.)
Στατιστικό z: $z = \frac{\hat{p_1}-1/2}{\sqrt{(\frac{1/2*(1-1/2))}{34}}} = 0.6859943$
Άρα το $pvalue$ ισούται με $1-\Phi(|z|) = 0.2463584$.Η τιμή του $pvalue$ είναι μεγάλη,δεν μπορούμε επομένως να απορρίψουμε την μηδενική υπόθεση.Άρα ο αριθμός των κόκκινων smarties δεν είναι στατιστικά σημαντικά μεγαλύτερος από των μπλε.

b.

Θα κάνουμε χ^2 έλεγχο καλής προσαρμογής:
H_0:Η κατανομή είναι ίδια με του 2009
H_α:Η κατανομή έχει αλλάξει

Δεδομένα:

Κόκκινο	19
Καφέ	22
Πράσινο	8
Κίτρινο	16
Μπλε	15
	80

Μέσο Πλήθος Παρατηρήσεων υπό την H_0:

Κόκκινο	14.24
Καφέ	15.84
Πράσινο	20.16
Κίτρινο	14.08
Μπλε	15.68
	80

$$x^2 = \frac{(19-14.24)^2}{14.24} + \frac{(22-15.84)^2}{15.84} + \frac{(8-20.16)^2}{20.16} + \frac{(16-14.08)^2}{14.08} + \frac{(15-15.68)^2}{15.68} = 11.61259$$

Υπολογισμός $pvalue$ στην R:
pchisq(x,df=4,lower.tail=FALSE)
0.02047712

Το $pvalue = 0.02047712$ έχει πολύ μικρή τιμή επομένως απορρίπτεται η μηδενική υπόθεση. Άρα η κατανομή έχει αλλάξει.

c.

Ελέγχουμε αν τα τυχαία δείγματα από smarties και M&M's έχουν την ίδια κατανομή χρωμάτων.

Θα κάνουμε χ^2 έλεγχο ομοιογένειας:
H_0:Η κατανομή των χρωμάτων είναι ίδια στα smarties και τα M&M's
H_α:Η κατανομή δεν είναι ίδια.

Πίνακας συνάφειας:

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

```
> data<-matrix(c(19,12,22,10,8,5,16,20,15,9),ncol=2,byrow=TRUE)
> colnames(data)<-c("smarties","M&M's")
> rownames(data)<-c("red","brown","green","yellow","blue")
> data<-as.table(data)
> data
       smarties M&M's
red          19    12
brown        22    10
green         8     5
yellow       16    20
blue         15     9
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 4.6262, df = 4, p-value = 0.3278
```

Όπως προκύπτει απο την R,$df = 4$, $\chi^2 = 4.6262$ και $pvalue = 0.3278$. Το $pvalue$ είναι μεγάλο,επομένως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση.Άρα οι κατανομές των χρωμάτων των smarties και των M&M's δεν έχουν στατιστικά σημαντική διαφορά.