ΣΤΑΤΙΣΤΙΚΗ ΣΤΗ ΠΛΗΡΟΦΟΡΙΚΗ

ΕΡΓΑΣΙΑ 3

ΜΕΛΗ ΟΜΑΔΑΣ:

ΑΘΑΝΑΣΙΟΣ ΚΛΕΤΤΑΣ: 3180079

ΙΑΣΩΝ ΖΙΩΓΑΣ: 3180057

ΑΣΚΗΣΗ 1:

a)

Γνωρίζουμε ότι κάνουμε 50 ρίψεις νομίσματος οπότε τα δεδομένα μας είναι κατάλληλα για την παρακάτω μεθοδολογία.

Για να υπολογίσουμε το 95% διάστημα εμπιστοσύνης για την συχνότητα της κορώνας χρησιμοποιούμε τον εξής τύπο :

$$\hat{p} \pm z_* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Όπου

- $\hat{p}$ = 29/50 = 0.58
- z_*= 1.96
- n = 50

Το 95% διάστημα εμπιστοσύνης που παίρνουμε είναι το :

(0.4431926, 0.7168074)

b)

Κάνουμε τις εξής υποθέσεις :

- H_0 : p = 0.5 ,
- H_α: p ≠ 0.5

όπου p το ποσοστό φορών που ήρθε το νόμισμα κορώνα.

Βρίσκουμε το z χρησιμοποιώντας το παρακάτω τύπο :

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Όπου

- $\hat{p}$ = 29/50 = 0.58
- p_0= 1.96
- n = 50

Η τιμή του z είναι 1.131371 επομένως μπορούμε να υπολογίσουμε το p-value.

p-value = 2Φ(-|z|) = 0.257899

Παρατηρούμε ότι p-value > α. Άρα δεν μπορούμε να απορρίψουμε την μηδενική μας υπόθεση.

c)

Για να βρούμε πόσες ρίψεις θα έπρεπε να πραγματοποιήσουμε εάν το περιθώριο λάθους στο διάστημα του ερωτήματος (a) θα θέλαμε να είναι μικρότερο του 1% θα πρέπει :

$$n \geq \frac{z_*^2 p(1-p)}{m^2}$$

Όπου

- p = 0.5
- z_*= 1.96
- m = 0.01

Άρα βρίσκουμε ότι θα πρέπει να είναι : n ≥ 9604 .

ΑΣΚΗΣΗ 2:

Παρατηρούμε ότι η ανισότητα :

$$n \geq \frac{z_*^2 p(1-p)}{m^2}$$

δεν επηρεάζεται από τον αριθμό του πληθυσμού αλλά από το διάστημα εμπιστοσύνης(C και κατ' επέκταση το z*) και το περιθώριο λάθους (m) γεγονός που σημαίνει ότι τόσο στην Ελλάδα όσο και στις ΗΠΑ θα χρησιμοποιηθούν 1100 άτομα στην δειγματοληψία.

ΑΣΚΗΣΗ 3:

a)

Ο αριθμός των δεδομένων μας είναι αρκετά μεγάλος (n=60) γεγονός το οποίο μας επιτρέπει να χρησιμοποιήσουμε την παρακάτω μεθοδολογία

Κάνουμε τις εξής υποθέσεις :

- $H_0 : p_1 = p_2$,
- $H_\alpha : p_1 \neq p_2$

Όπου p_1 το ποσοστό των ανδρών που καπνίζουν και p_2 το ποσοστό των γυναικών που καπνίζουν.

Αρχικά χρησιμοποιούμε τον τύπο :

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

Όπου

- $\hat{p}_1 = 12 / 30 = 0.4$
- $\hat{p}_2 = 14 / 30 = 0.4666667$
- $\hat{p} = 26 / 60 = 0.4333333$
- $n_1 = 30$
- $n_2 = 30$

Η τιμή του z είναι -0.5210501επομένως μπορούμε να υπολογίσουμε το p-value.

$$\text{p-value} = 2\Phi(-|z|) = 0.6023319$$

Άρα δεν μπορούμε να απορρίψουμε την μηδενική μας υπόθεση.

b)

Για να υπολογίσουμε 95% διάστημα εμπιστοσύνης για τη διαφορά του ποσοστού καπνιστών μεταξύ ανδρών και γυναικών χρησιμοποιούμε τον τύπο :

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

Όπου

- $\hat{p}_1$ = 12 / 30 = 0.4
- $\hat{p}_2$ = 14 / 30 = 0.4666667
- n_1 = 30
- n_2 = 30

Το 95% διάστημα εμπιστοσύνης που παίρνουμε είναι το :

(-0.3168743, 0.1835410)

c)

Με την βοήθεια της R και της συνάρτησης chisq.test()

X^2 = 0.27149 και p-value = 0.6023

Ο πίνακας συνάφειας βάσει των δεδομένων του Πίνακα 1 είναι ο εξής :

	MALE	FEMALE	
SMOKER	12	14	26
NOT SMOKER	18	16	34
	30	30	60

d)

Παρατηρούμε ότι το p-value του X^2 ελέγχου είναι ίδιο με το p-value που βρήκαμε από το (a) ερώτημα και αυτό είναι λογικό καθώς γνωρίζουμε ότι ο δίπλευρος έλεγχος είναι ισοδύναμος με τον X^2 έλεγχο σε πίνακες συνάφειας 2x2.

ΑΣΚΗΣΗ 4:

a)

Ο αριθμός των δεδομένων μας είναι αρκετά μεγάλος (n=80) γεγονός το οποίο μας επιτρέπει να χρησιμοποιήσουμε την εκάστοτε μεθοδολογία

Αφού τα δεδομένα μας αποτελούν μέρος ενός πληθυσμού (σακουλάκι smarties) θα χρησιμοποιήσουμε τον τύπο:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Όπου για $\hat{p}$ =μέση τιμή των κόκκινων smarties και για p_0 = μέση τιμή των μπλε smarties

- $\hat{p}$ = 0.2375
- p_0= 0.1875
- $n = 80$
- H_0: $\hat{p}$ = p0
- H_a: $\hat{p}$ > p0

Το z = 1.145784

p-value = 0.12594323

```
> red_mean <- mean(data$color=="red")
> blue_mean <- mean(data$color == "blue")
> red_mean
[1] 0.2375
> blue_mean
[1] 0.1875
> #H0 red_mean = blue_mean , #Ha red_mean > blue_mean
> z <- (red_mean-blue_mean)/sqrt((blue_mean*(1-blue_mean))/80)
> z
[1] 1.145784
> 1-pnorm(z)
[1] 0.1259423
```

Αφού το p-value είναι αρκετά μεγάλο δεν μπορούμε να απορρίψουμε την αρχική μας υπόθεση.

b)

Αφού θα χρειαστεί να ελέγξουμε εάν μια κατηγορική μεταβλητή επηρεάζει μια άλλη κατηγορική θα χρησιμοποιήσουμε έναν έλεγχο X^2 μέσω ενός πίνακα

```
> t

          2009  2020
  blue   19.60 18.75
  brown  19.80 27.50
  green  25.20 10.00
  red    17.80 24.75
  yellow 17.60 20.00
> chisq.test(t,correct=FALSE)

        Pearson's Chi-squared test

data:  t
X-squared = 9.1196, df = 4, p-value = 0.05818
```

Το p-value μας δείχνει την πιθανότητα να πάρουμε τα συγκεκριμένα δεδομένα εάν η κατηγορική μεταβλητή χρόνος δεν ήταν επεξηγηματική. Με p-value = 0.05818 εάν και δεν γνωρίζουμε το επίπεδο σημαντικότητας με βάση την κρίση μας καταλήξαμε ότι η μεταβλητή χρόνος επηρεάζει τον αριθμό χρωμάτων των smarties.

c)

Παρόμοια με την b θα χρειαστεί να ελέγξουμε εάν μια κατηγορική μεταβλητή (σε αυτήν την περίπτωση η εταιρία παραγωγής) επηρεάζει μια άλλη κατηγορική μεταβλητή(ο αριθμός των χρωμάτων των κουφέτων). Έτσι θα χρησιμοποιήσουμε και σε αυτήν την περίπτωση έναν έλεγχο X^2 μέσω ενός πίνακα.

```
> c

        m&ms smarties
  blue     9       15
  brown   10       22
  green    5        8
  red     12       19
  yellow  20       16
> chisq.test(c,correct=FALSE)

        Pearson's Chi-squared test

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

Το p-value μας δείχνει την πιθανότητα να πάρουμε τα συγκεκριμένα δεδομένα εάν η κατηγορική μεταβλητή εταιρία παραγωγής δεν ήταν επεξηγηματική(δεν επηρέαζε την μεταβλητή αριθμός κουφέτων ανά χρώμα). Με p-value = 0.3278 εάν και δεν γνωρίζουμε το επίπεδο σημαντικότητας με βάση την κρίση μας καταλήξαμε ότι η μεταβλητή εταιρία παραγωγής δεν επηρεάζει τον αριθμό των κουφέτων κάθε χρώματος.