Στατιστική Στην Πληροφορική - $3^{η}$ Σειρά Ασκήσεων
ΟΠΑ , Ακαδημαϊκό Έτος: 2020-2021

Ομάδα #0

- ΤΣΙΟΜΠΙΚΑΣ ΔΗΜΗΤΡΙΟΣ , 3180223
- ΠΑΝΑΓΙΩΤΟΥ ΠΑΝΑΓΙΩΤΗΣ , 3180139

1η Άσκηση

a. Αρχικά βρίσκουμε το z ως εξής :

```
[1] 1.644854
> z <- qnorm(0.975)
> z
[1] 1.959964
```

Και χρησιμοποιούμε αυτόν τον τύπο ώστε να βρούμε το Διάστημα Εμπιστοσύνης 95%

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

Έχουμε :

n = 50

X = 29

p = X/n = 29/50 = 0.58

z = 1.959964

```
> p = x/n
> p + c(-1,1) * z * sqrt((p * (1-p)) / n)
[1] 0.4431951 0.7168049
```

Άρα το Διάστημα εμπιστοσύνης 95% για τη συχνότητα εμφάνισης της κορώνας είναι **[0.4431951,0.7168049].**

b. Έχουμε α = 0.05

H0 : p = ποσοστό εμφάνισης κορώνας-γραμμάτων ισούται με 50% (Για να είναι δίκαιο το νόμισμα πρέπει να έχουμε ίσο αριθμό εμφάνισης κορώνας-γραμμάτων)

Ha : p <> 50%

Για να βρούμε το z χρησιμοποιούμε τον τύπο :

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

p = 0.58

p0 = 0.5

n = 50

```
> p0 <- 0.5
> z <- (p - p0) / (sqrt((p0 * (1-p0)) / n))
> z
[1] 1.131371
```

z = 1.131371

άρα pvalue = 2Φ(-|z|) = 2 * 0.1292 = 0.2584

Καταλήγουμε λοιπόν στο συμπέρασμα ότι pvalue > α άρα δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση , το οποίο σημαίνει ότι το νόμισμα είναι δίκαιο παρά τις περισσότερες εμφανίσεις κορώνας στο δείγμα μας.

c. Έχουμε m = 0.01

Χρησιμοποιούμε τον τύπο :

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

Χρησιμοποιώντας τα παραπάνω δεδομένα (με z = 1.96) παίρνουμε το αποτέλεσμα :

```
> (z ^ 2 * p * (1-p)) / m ^ 2
[1] 9357.794
```

Άρα πρέπει να κάνουμε n ≃ 9358 ρίψεις ή παραπάνω με το νόμισμα για να έχουμε περιθώριο λάθους μικρότερο του 1%.

2η Άσκηση

Έχουμε:

m = 0.03

z = 1.959964 (επειδή έχουμε 95% Διάστημα Εμπιστοσύνης)

Εδώ δεν γνωρίζουμε το p οπότε θα χρησιμοποιήσουμε τον τύπο :

$$n \geq \frac{z_*^2}{4m^2}$$

```
> (z ^ 2) / (4 * m ^ 2)
[1] 1067.072
```

Άρα θα έπρεπε να πάρουν n ≃ 1067 ή παραπάνω δείγματα για να έχουν περιθώριο λάθους 3%. Παρατηρούμε πως ο γενικός πληθυσμός δεν έχει τόσο μεγάλη σημασία καθώς όταν θέλουμε να βγάλουμε ένα συμπέρασμα με τις δοσμένες συνθήκες (95% ΔΕ και 3% περιθώριο λάθους) αρκεί να πάρουμε 1067 ή παραπάνω δείγματα.

3η Άσκηση

a. Αρχικά θα πάρουμε τα ποσοστά των ανδρών και γυναικών που καπνίζουν ώστε να δούμε αν σχετίζεται το φύλο με το κάπνισμα.
Έχουμε :
H0 : p1 = p2 (δεν έχει σχέση το φύλο με το κάπνισμα)
Ha : p1 <> p2 (έχει σχέση το φύλο με το κάπνισμα)
nFem = 30
nMale = 30
pFem = 14/30 = 0.4666667

pMale = 0.4

z = 0.5210501

```
> z <- (pFem - pMale) / (sqrt((p * (1-p) / nFem) + (p * (1-p)/nMa
le)))
```

pvalue = 2Φ(|-z|) = 2 * 0.3015 = 0.603

Άρα pvalue > α (μεγαλύτερο από κάθε α) οπότε δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση το οποίο σημαίνει ότι δεν μπορούμε να συμπεράνουμε αν το φύλο έχει τόσο σχέση με το κάπνισμα.

b. Θα χρησιμοποιήσουμε τον τύπο :

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

```
> pFem - pMale + c(-1,1) * z * sqrt(((pFem * (1-pFem) / nFem) +
  (pMale * (1-pMale) / nMale)))
[1] 0.0001510006 0.1331823328
```

Άρα το Διάστημα Εμπιστοσύνης 95% είναι

[0.00015100006 , 0.1331823328].

c.

	Γυναίκες	Άνδρες	
Μη-Καπνιστές	16	18	34
Καπνιστές	14	12	26
	30	30	60

	Γυναίκες	Άνδρες	
Μη-Καπνιστές	(34*30)/60 = 17	(34*30)/60 = 17	34
Καπνιστές	(26*30)/60 = 13	(26*30)/60 = 13	26
	30	30	60

$X^2 = ((16-17)^2 / 17) + ((18-17)^2 / 17) + ((14-13)^2 / 13) + ((12-13)^2 / 13) = 0.2714932$

$df = (r-1) * (c-1) = (3-1) * (3-1) = 4$

d. Παίρνοντας α = 0.05 με df = 4 έχουμε Pvalue = 14.860 > α (απαντήσαμε βάσει του πίνακα που δόθηκε στην εκφώνηση) άρα συμπεραίνουμε και πάλι ότι δεν υπάρχει συσχέτιση ανάμεσα στο φύλο και το κάπνισμα. Παρατηρούμε ότι το pvalue της X^2 και του ελέγχου z δίνουν το ίδιο συμπέρασμα παρά τη διαφορά των ποσοστών των pvalues (X^2 pvalue ≃ 15% και z pvalue ≃ 60%).

4η Άσκηση

a. Θα χρησιμοποιήσουμε τον έλεγχο goodness of fit.
Έχουμε :
H0 : έχουμε ίδια ποσότητα κόκκινων και μπλε smarties
Ha : Έχουμε μεγαλύτερη ποσότητα κόκκινων από ότι μπλε smarties
Φτιάχνουμε τους πίνακες :

	Oi
Κόκκινα	19
Μπλε	15
	34

Για τον πίνακα E έχουμε ότι η μέση τιμή εφόσον έχουμε 2 κατηγορίες smarties θα είναι : ½ * 34 = 17

Άρα :

	Ei
Κόκκινα	17
Μπλε	17
	34

Για το X^2 στατιστικό ελέγχου έχουμε :

```
> xSquare <- (19-17)^2/17 + (15-17)^2/17
> xSquare
[1] 0.4705882
```

Άρα **X^2 = 0.4705882**

Df = (r-1) = 1

```
> 1 - pchisq(xSquare,1)
[1] 0.4927167
```

Άρα το pvalue = 0.4927167 > α άρα δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση το οποίο σημαίνει ότι έχουμε ίδια ποσότητα κόκκινων και μπλε smarties.

b.

Έχουμε :

H0 : δεν έχει αλλάξει η κατανομή από το 2009

Ha : Έχει αλλάξει η κατανομή από το 2009

α = 0.05 (5%)

Τοποθετούμε τα στοιχεία του 2009 σε πίνακα :

```
> color <- c(replicate(22,"Brown"),replicate(19,"Red"),replicate
(16,"Yellow"),replicate(15,"Blue"),replicate(8,"Green"))
> data <- table(color)
> data
color
  Blue  Brown  Green    Red Yellow
    15     22      8     19     16
```

Και εκτελούμε X^2 έλεγχο με τα δοσμένα ποσοστά :

```
> chisq.test(data,p = c(0.196,0.198,0.252,0.178,0.176))

        Chi-squared test for given probabilities

data:  data
X-squared = 11.613, df = 4, p-value = 0.02048
```

Παρατηρούμε ότι το pvalue = 0.02% < α άρα αποκλείουμε την μηδενική υπόθεση και πάμε στην εναλλακτική. Αυτό μας οδηγεί στο συμπέρασμα ότι η κατανομή έχει αλλάξει από το 2009.

c.

Έχουμε τον εξής έλεγχο υπόθεσης :

H0 : η αναλογία στα smarties και τα M&Ms είναι η ίδια

Ha : η αναλογία στα smarties και τα M&Ms δεν είναι η ίδια

Έχουμε τον εξής πίνακα :

```
> percentagesMat
       Smarties M&Ms
Brown        22   10
Red          19   12
Yellow       16   20
Blue         15    9
Green         8    5
```

Βρίσκουμε τα ποσοστά :

```
> prop.table(percentagesMat,2)
       Smarties        M&Ms
Brown    0.2750 0.17857143
Red      0.2375 0.21428571
Yellow   0.2000 0.35714286
Blue     0.1875 0.16071429
Green    0.1000 0.08928571
```

Παρατηρούμε ότι τα ποσοστά δεν είναι αντιπροσωπευτικά ώστε να βγάλουμε συμπέρασμα και εκτελούμε X^2 έλεγχο στον πίνακα :

```
> chisq.test(percentagesMat)

        Pearson's Chi-squared test

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

Παρατηρούμε ότι pvalue = 0.3278 > α οπότε δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση το οποίο σημαίνει ότι η αναλογία στα smarties και τα M&Ms είναι ίδια.