ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Στατιστική στην Πληροφορική (2022-2023)

Διονύσιος Ρηγάτος (P3200262)

Εργασία 3

Άσκηση 1

Α) Θα πάρουμε:

- # Ρίψεων με αποτέλεσμα κορώνα (29)
- # Ρίψεων με αποτέλεσμα γράμματα (21)

Έχουμε $29 > 15$ & $21 > 15$ επομένως μπορούμε να βγάλουμε διάστημα εμπιστοσύνης με ακρίβεια.

Άρα : $\hat{\boldsymbol{p}} = 29/50 = 0.58$

Και συνεπώς:

```
95% Confidence Interval: [0.443195075474522, 0.716804924525478]
```

Β) Θα πάρουμε ως μηδενική υπόθεση την περίπτωση που το νόμισμα είναι δίκαιο, συνεπώς έχει πιθανότητα ½ κάθε ρίψη να είναι είτε κορώνα είτε γράμματα. Έχουμε μέσο πλήθος επιτυχιών/αποτυχιών ($n = n(p\text{-}1/2) = 1/2$) ίσο με $50*1/2 = 25 > 10$, άρα τα αποτελέσματά μας θα είναι

Ho: $p = ½$ (Δίκαιο νόμισμα)

Ha: $p \neq ½$ (Μη δίκαιο νόμισμα)

```
z: 1.13137084989848
p-value: 0.25789903529234
```

Έχουμε 25.7% = p-value > α = 5% και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται. Άρα το νόμισμα είναι δίκαιο.

Γ)

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

Όπου:

```
res <- (qnorm(0.975)^2) / (4*(0.01)^2)
```

```
Number of throws (rounded): 9604
```

Άρα θα πρέπει να ρίξουμε το νόμισμα 9604 φορές έτσι ώστε για confidence interval 95% να έχουμε margin of error < 1%.

Άσκηση 2

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

Όπου:

```
res <- (qnorm(0.975)^2) / (4*(0.01)^2)
```

```
Sample required (people - rounded): 1068
```

Το μέγεθος του δείγματος των δημοσκοπήσεων δεν εξαρτάται από το μέγεθος του πληθυσμού αλλά από τα z^* και m, επομένως ήταν αναμενόμενο ότι τα 1100 άτομα θα ήταν αρκετά.

Άσκηση 3

Α) Τα δεδομένα αποτελούν SRS και είναι περισσότερα του 5 σε κάθε περίπτωση, επομένως αποτελούν καλά δεδομένα για να εφαρμόσουμε έλεγχο σημαντικότητας.

```
# Male Smokers: 12
# Male Non-Smokers: 18
# Female Smokers: 14
# Female Non-Smokers: 16
```

Έστω pM και pF οι πιθανότητες να καπνίζει ένας άντρας και μία γυναίκα respectively. Θέλουμε να εξετάσουμε εάν υπάρχει σχέση μεταξύ φύλου και καπνίσματος.

Ho: pM = pF

Ha: pM ≠ pF

```
Probability a Male is a smoker: 0.4
Probability a Female is a smoker: 0.466666666666667
Probability a Person is a smoker: 0.433333333333333
```

```
z: 0.521050105718906
p-value: 0.602331867061728
```

60% = p-value > α = 5% (διάστημα εμπιστοσύνης 95%) και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται και άρα δεν υπάρχει σχέση μεταξύ φύλου και καπνίσματος.

Β) Όπως προαναφέρθηκε, τα δείγματα αποτελούν SRS και είναι αρκετά (περισσότερα του 10) και συνεπώς μπορούμε να υπολογίσουμε ακριβές διάστημα εμπιστοσύνης.

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

```
95% Confidence Interval: [-0.183536353203303, 0.316869686536636]
```

Γ) Γνωρίζουμε ότι τα δείγματά μας λήφθηκαν με τυχαία δειγματοληψία. Θα εξετάσουμε εάν είναι ανεξάρτητα.

Ho: Το φύλο και το κάπνισμα είναι ανεξάρτητα μεταξύ τους.

Ha: Το φύλο και το κάπνισμα δεν είναι ανεξάρτητα μεταξύ τους.

Ακολουθεί ο πίνακας συνάφειας.

	SEX		
SMOKER	F	M	Sum
N	16	18	34
Y	14	12	26
Sum	30	30	60

Δ)

```
  Pearson's Chi-squared test

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

p-value = 0.6023 που ισούται με το αποτέλεσμα του ερωτήματος (Α). Άρα το χ^2 test ισούται με το αποτέλεσμα του z test.

Τέλος παρατηρούμε ότι το χ^2 αποτελεί την τετραγωνική ρίζα του z.

```
z: 0.521050105718906
```

Άσκηση 4

Σε όλη την άσκηση έχουμε απλό τυχαίο δείγμα και αρκετά δείγματα (>10). Επίσης, είναι σημαντικό να τονίσουμε ότι η εντολή της R για chi squared παράγει warnings σε περίπτωση που τα δεδομένα δεν είναι κατάλληλα για τον έλεγχο.

Α)

Ho: pK <= ½ (Δεν παρασκευάζονται περισσότερα κόκκινα smarties απ' ότι μπλε)

Ha: pK > ½ (Παρασκευάζονται περισσότερα κόκκινα smarties απ' ότι μπλέ)

Θα χρησιμοποιήσουμε το εργαλείο prop.test της R (ουσιαστικά z έλεγχος για 1 δείγμα), το οποίο θα ελέγξει τα proportions για το 1 δείγμα που έχουμε στην διάθεσή μας και θα μας δώσει το p-value έτσι ώστε να απορρίψουμε η όχι την μηδενική υπόθεση.

```
    1-sample proportions test without continuity correction

data:  19 out of 34, null probability 0.5
X-squared = 0.47059, df = 1, p-value = 0.2464
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4196133 1.0000000
sample estimates:
        p
0.5588235
```

Το αποτέλεσμα είναι πως p-value $= 0.2464$ και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται, και άρα δεν παράγονται περισσότερα κόκκινα smarties απ' ότι μπλέ.

Β)

Ho: Η κατανομή του πληθυσμού των χρωμάτων συμφωνεί με καφέ, κόκκινο, κίτρινο, μπλε και πράσινο, είναι 19.8%, 17.8%, 17.6%, 19.6% και 25.2% αντίστοιχα.

Ha: Η κατανομή του πληθυσμού των χρωμάτων διαφέρει από τις παραπάνω τιμές.

```
data <- c(22, 19, 16, 15, 8)
```

```
      Chi-squared test for given probabilities

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

Άρα p-value $= 0.02048 < \alpha = 0.05$ (για διάστημα εμπιστοσύνης 95%) είναι σημαντικά μικρό και συνεπώς η μηδενική υπόθεση απορρίπτεται. Άρα η κατανομή έχει αλλάξει από το 2009.

Γ) Θα χρησιμοποιήσουμε χ^2 έλεγχο για ομοιογένεια

Ho: Smarties και M&Ms έχουν την ίδια αναλογία χρωμάτων.

Ha: Οι αναλογίες διαφέρουν.

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

```
        Pearson's Chi-squared test

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

Έχουμε p-value $= 0.3278 > \alpha = 0.05$ και συνεπώς η μηδενική υπόθεση δεν απορρίπτεται, άρα τα Smarties και τα M&Ms έχουν την ίδια αναλογία χρωμάτων.

Κώδικας R

[Google Drive Link]

Ο κώδικας δεν γράφτηκε για να βαθμολογηθεί και συνεπώς είναι προχειρογραμμένος, χωρίς comments και μπορεί να περιέχει περιττές (η να λείπουν) και ακόμα και λανθασμένες εντολές και συμπεριλαμβάνεται μόνο για πληρότητα και ακεραιότητα ώστε οι πράξεις να μην φανούν ουρανοκατέβατες.