Χρήστος Πατρινόπουλος
3200150
Αλβιόνα Μάντσο
3200098

Στατιστική στην Πληροφορική

Άσκηση 1

a) Το επίπεδο εμπιστοσύνης είναι 95%, οπότε για να είναι ακριβή τα αποτελέσματα χρειαζόμαστε:

- # «επιτυχιών» (κορώνες) = 29 > 15 ✓
- # «αποτυχιών» (γράμματα) = 21 > 15 ✓

$$\hat{p} = \frac{29}{50} = 0.58 = \mathbf{58\%}$$

```
{ p_hat + c(-1, 1)*(-1)*qnorm(0.025)*sqrt(p_hat*(1-p_hat)/n) }
```

Το διάστημα εμπιστοσύνης για επίπεδο εμπιστοσύνης 95% είναι **[0.4431951, 0.7168049].**

b) $H_o : p = 0.5$ «Το νόμισμα είναι δίκαιο»
$H_a : p \neq 0.5$ «Το νόμισμα δεν είναι δίκαιο»

Για να είναι ακριβή τα αποτελέσματα του ελέγχου σημαντικότητας χρειαζόμαστε:

- Μέσο πλήθος «επιτυχιών» (κορώνες)
  = n* 0.5 = 50*0.5 =25 >10 ✓
- Μέσο πλήθος «αποτυχιών» (γράμματα)
  = n* (1 - 0.5) = 50*0.5 =25 >10 ✓

z = `{ (p_hat - 0.5)/sqrt(0.5*0.5/n) }` = **1.131371**
p-value = `{ 2*pnorm(-z) }` = **0.257899** ~ 26%

Με επίπεδο σημαντικότητας 5%, το p-value είναι ~26% **>** 5% και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση ($p = 0.5$). Συνεπώς τα δεδομένα δεν είναι στατιστικά σημαντικά για την

εναλλακτική υπόθεση ($p \neq 0.5$). Οπότε συμπεραίνουμε ότι το νόμισμα είναι δίκαιο.

c) Με χρήση του τύπου:

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

λαμβάνουμε:

$$n \geq \frac{1.96*1.96}{4*0.01*0.01} = 9604$$

Δηλαδή θα πρέπει να εκτελέσουμε 9604 ρίψεις για διάστημα εμπιστοσύνης 95% με περιθώριο λάθους μικρότερο 1%.

Άσκηση 2

$C\% = 95\%$, $m = 3\%$

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

Με βάση τον τύπο και για δεδομένο επίπεδο εμπιστοσύνης (άρα προκαθορισμένο z^*) και περιθώριο σφάλματος m, το μέγεθος του δείγματος που απαιτείται δεν εξαρτάται από το μέγεθος του πραγματικού πληθυσμού, παρά μόνο από τα z^* και m. Συνεπώς, θα προκύψει ότι και στις ΗΠΑ το ελάχιστο δείγμα που χρειαζόμαστε είναι 1100 άτομα.

Πράγματι, αν επιχειρούσαμε να κάνουμε τους υπολογισμούς θα παίρναμε:

$$n \geq \frac{1.96*1.96}{4*0.03*0.03} = 1067.111 \sim 1068$$

δλδ περίπου 1100 άτομα στο δείγμα (όπως και στην Ελλάδα).

Άσκηση 3

a. Για να είναι ακριβής ο έλεγχος σημαντικότητας θα πρέπει:
   - Τα δείγματα να έχουν ληφθεί με ανεξάρτητο τρόπο από τους 2 πληθυσμούς (άντρες, γυναίκες) και το καθένα να είναι SRS.
   - Για τις γυναίκες:
     # «επιτυχιών» (καπνίστριες) = 14 > 5
     # «αποτυχιών» (μη καπνίστριες) = 16 > 5

     Για τους άντρες:
     # «επιτυχιών» (καπνιστές) = 12 > 5
     # «αποτυχιών» (μη καπνιστές) = 18 > 5

$H_o : p_f = p_m \Rightarrow p_f - p_m = 0$
$H_a : p_f \neq p_m \Rightarrow p_f - p_m \neq 0$

$n_f =$ `sum(SEX=='F')`
$\hat{p}_f =$ { `sum(SEX=='F' & SMOKER=='Y')/n_f` } $= 0.4666667$

$n_m =$ `sum(SEX=='M')`
$\hat{p}_m =$ { `sum(SEX=='M' & SMOKER=='Y')/n_m` } $= 0.4$

$p =$ { `sum(SMOKER=='Y'))/(n_f + n_m)` } $= 0.4333333$

$z =$ {`(pf_hat-pm_hat)/sqrt(p*(1-p)/n_f + p*(1-p)/n_m`}
$= 0.5210501$

p-value = { `2*pnorm(-z)` } = 0.6023319

Το p-value είναι ~60% και συνεπώς αρκετά μεγάλο ώστε δεν απορρίπτουμε τη μηδενική υπόθεση ($p_f = p_m$). Συνεπώς τα δεδομένα δεν είναι στατιστικά σημαντικά για την εναλλακτική υπόθεση ($p_f \neq p_m$). Οπότε συμπεραίνουμε ότι δεν υπάρχει σχέση μεταξύ φύλου και καπνίσματος (δλδ το φύλο δεν επηρεάζει το κάπνισμα).

b. Για να είναι ακριβής ο υπολογισμός του διαστήματος εμπιστοσύνης θα πρέπει:
   - Τα δείγματα να έχουν ληφθεί με ανεξάρτητο τρόπο από τους 2 πληθυσμούς (άντρες, γυναίκες) και το καθένα να είναι SRS. ✓
   - Για τις γυναίκες:
     # «επιτυχιών» (καπνίστριες) = 14 > 10 ✓
     # «αποτυχιών» (μη καπνίστριες) = 16 > 10 ✓

     Για τους άντρες:
     # «επιτυχιών» (καπνιστές) = 12 > 10 ✓
     # «αποτυχιών» (μη καπνιστές) = 18 > 10 ✓

```
{(pf_hat - pm_hat) + c(-1,1)*(-1) *qnorm(0.025)*
sqrt(pf_hat*(1-pf_hat)/sum(SEX == 'F') + pm_hat*(1-
pm_hat)/sum(SEX == 'M'))}
```

Το διάστημα εμπιστοσύνης για επίπεδο εμπιστοσύνης 95% είναι **[-0.1835364, 0.3168697].**

c. Έλεγχος ανεξαρτησίας

H_o : Το φύλο και το κάπνισμα είναι ανεξάρτητα
H_a : Το φύλο και το κάπνισμα δεν είναι ανεξάρτητα

Πίνακας συνάφειας:

```
> addmargins(table(SMOKER, SEX))
      SEX
SMOKER  F  M Sum
   N   16 18  34
   Y   14 12  26
   Sum 30 30  60
```

d. Σημείωση: *Τα δεδομένα προέρχονται από SRS και η εκτέλεση της εντολής παρακάτω δεν παρήγαγε κάποιο warning οπότε τα δεδομένα που δόθηκαν ήταν κατάλληλα για την εφαρμογή του ελέγχου.*

```
> t = table(SMOKER, SEX)
> chisq.test(t, correct=FALSE)

        Pearson's Chi-squared test

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

- Παρατηρούμε ότι το p-value που προκύπτει με τον x^2 έλεγχο είναι **0.6023** δλδ **ίσο** με αυτό που υπολογίσαμε στο ερώτημα (α) με χρήση του z ελέγχου. Αυτό ήταν αναμενόμενο εφόσον ο πίνακας συνάφειας είναι διαστάσεων 2x2, περίπτωση στην οποία μπορεί να εφαρμοστεί z έλεγχος.
- Επίσης παρατηρούμε ότι x^2-στατιστικό = 0.27149 = (z-στατιστικό)^2
  [αφού (z-στατιστικό)^2 = 0.5210501^2 = 0.27149].

Άσκηση 4

a. Για να είναι ακριβής ο υπολογισμός θα πρέπει:
   - Το δείγμα να έχει ληφθεί με SRS από τον πληθυσμό (όπου εδώ πληθυσμός είναι τα κόκκινα και μπλε smarties που παρασκευάζονται και δείγμα τα 34 συνολικά smarties της σακούλας που είναι είτε μπλε είτε κόκκινα)
   - Μέσο πλήθος επιτυχιών $n(p_0) = 34*0.5 = 17 > 10$
   - Μέσο πλήθος αποτυχιών $n(1-p_0) = 34*0.5 = 17 > 10$

$H_o : p_{red} \leq 0.5$ «Δεν παρασκευάζονται περισσότερα κόκκινα από μπλε»
$H_a : p_{red} > 0.5$ «Παρασκευάζονται περισσότερα κόκκινα από μπλε»

$\hat{p}_{red}$ = 19 / (19+15) = 19/ 34 = 0.5588235

$z = (\hat{p}_{red} - 0.5) / \sqrt{0.5(1-0.5)/34} = 0.68599434$

$p - value = \{ 1 - \text{pnorm}(z) \} = 0.2463583$

Το p-value είναι ~25% και συνεπώς αρκετά μεγάλο ώστε δεν απορρίπτουμε τη μηδενική υπόθεση ($p_{red} \leq 0.5$). Συνεπώς τα δεδομένα δεν είναι στατιστικά σημαντικά για την εναλλακτική υπόθεση ($p_{red} > 0.5$). Οπότε συμπεραίνουμε ότι δεν παρασκευάζονται περισσότερα κόκκινα smarties από ότι μπλε.

b.

H_o : Η κατανομή του πληθυσμού συμφωνεί με την

p = <0.198, 0.178, 0.176, 0.196, 0.252>

H_a : Ο πληθυσμός έχει διαφορετική κατανομή

```
> d <- c(22,19,16,15,8)
> chisq.test(d, p=c(0.198, 0.178, 0.176, 0.196, 0.252))

        Chi-squared test for given probabilities

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

Σημείωση: *Τα δεδομένα προέρχονται από SRS και η εκτέλεση της εντολής δεν παρήγαγε κάποιο warning οπότε τα δεδομένα που δόθηκαν ήταν κατάλληλα για την εφαρμογή του ελέγχου.*

Το p-value είναι ~2% και το θεωρούμε αρκετά μικρό ώστε απορρίπτουμε τη μηδενική υπόθεση. Συνεπώς τα δεδομένα είναι στατιστικά σημαντικά για την εναλλακτική υπόθεση. Οπότε συμπεραίνουμε ότι έχει αλλάξει η κατανομή από το 2009.

c. Έλεγχος ομοιογένειας

H_o : Η αναλογία χρωμάτων στα smarties είναι ίδια με αυτή στα M&Ms

H_a : Οι δύο αναλογίες είναι διαφορετικές

```
> t = matrix(c(22, 19, 16, 15, 8, 10, 12, 20, 9, 5), 5, 2)
> colnames(t) <- c('Smarties', 'M&Ms')
> rownames(t) <- c('Brown', 'Red', 'Yellow', 'Blue', 'Green')
> t <- as.table(t)
> t
       Smarties M&Ms
Brown        22   10
Red          19   12
Yellow       16   20
Blue         15    9
Green         8    5
> chisq.test(t, correct=FALSE)

        Pearson's Chi-squared test

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

Σημείωση: *Τα δεδομένα προέρχονται από SRS και η εκτέλεση της εντολής δεν παρήγαγε κάποιο warning οπότε τα δεδομένα που δόθηκαν ήταν κατάλληλα για την εφαρμογή του ελέγχου.*

Το p-value είναι ~33% και άρα αρκετά μεγάλο ώστε δεν απορρίπτουμε τη μηδενική υπόθεση. Συνεπώς τα δεδομένα δεν είναι στατιστικά σημαντικά για την εναλλακτική υπόθεση. Οπότε συμπεραίνουμε ότι η αναλογία χρωμάτων πρέπει να είναι η ίδια.