ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Συμπεράσματα για Αναλογίες και Πίνακες Κατηγοριών

Εργασία 3

Μαρία Σχοινάκη

Σοφία Παπαϊωάννου

Άσκηση 1

a) Το επίπεδο εμπιστοσύνης είναι 95%, οπότε:

Συνθήκες για την ακρίβεια του z-ελέγχου:

κορώνες = 29 > 15

γράμματα = 21 > 15

Οι συνθήκες αυτές ικανοποιούνται, συνεπώς μπορούμε να χρησιμοποιήσουμε την κανονική προσέγγιση.

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

Χρησιμοποιούμε το τύπο για το διάστημα εμπιστοσύνης της δυαδικής πιθανότητας:

$$CI = \hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ή αλγοριθμικά 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) H0 : p = 0.5 (Το νόμισμα είναι δίκαιο)
Ha : p = 0.5 (Το νόμισμα **δεν** είναι δίκαιο)

Συνθήκες για την ακρίβεια του z-ελέγχου:

1. Ο μέσος αριθμός κορώνων: $n \cdot p_0 = 50 \cdot 0.5 = 25 > 10$
2. Ο μέσος αριθμός γραμμάτων: $n \cdot (1-p_0) = 50 \cdot 0.5 = 25 > 10$

Οι συνθήκες αυτές ικανοποιούνται, συνεπώς μπορούμε να χρησιμοποιήσουμε την κανονική προσέγγιση.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \qquad p\text{-value} = 2 \cdot P(Z > |z|)$$

ή αλγοριθμικά

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≠0.5). Άρα καταλήγουμε ότι το νόμισμα μπορεί να θεωρηθεί **δίκαιο**.

c) Με χρήση του τύπου:

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

Παίρνουμε:

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

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

<u>Άσκηση 2</u>

- Επίπεδο εμπιστοσύνης: **95%**
- Περιθώριο σφάλματος: **3%** (m=0.03)

Ο τύπος για τον υπολογισμό του μεγέθους δείγματος είναι:

$$n \geq \frac{z^2}{4m^2}$$

Όπου:

- Z = 1.96: η τιμή z∗ για 95% επίπεδο εμπιστοσύνης,
- m: το περιθώριο σφάλματος.

Αντικαθιστούμε τις τιμές στον τύπο:

$$n \geq \frac{1.96^2}{4 \cdot 0.03^2}$$

= **1067.111 ≈ 1068**

Στρογγυλοποιούμε προς τα πάνω:

n ≈ 1100 άτομα

Παρατήρηση:

Το μέγεθος δείγματος **δεν εξαρτάται από τον πληθυσμό** όταν αυτός είναι πολύ μεγάλος (π.χ., Ελλάδα: 10 εκατομμύρια, Η.Π.Α.: 300 εκατομμύρια). Το μόνο που επηρεάζει το μέγεθος δείγματος είναι το επίπεδο εμπιστοσύνης (z) και το περιθώριο σφάλματος (m). Συνεπώς, το δείγμα στις Η.Π.Α. παραμένει ίδιο με αυτό της Ελλάδας, περίπου **1100 άτομα**.

Άσκηση 3

a) Για να είναι ακριβής ο έλεγχος σημαντικότητας θα πρέπει:

- Τα δείγματα να έχουν ληφθεί με ανεξάρτητο τρόπο από τους 2 πληθυσμούς 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

Ho : pf = pm => pf – pm = 0

Ha : pf ≠ pm => pf – pm ≠ 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 περίπου **60%**, πολύ μεγαλύτερο από το επίπεδο σημαντικότητας α=5. Εφόσον p-value > α, **δεν** απορρίπτουμε τη μηδενική υπόθεση (*H0 : pf=pm*).

Συνεπώς, τα δεδομένα **δεν** δείχνουν στατιστικά σημαντική διαφορά στα ποσοστά καπνιστών μεταξύ ανδρών και γυναικών. Το φύλο **δεν** φαίνεται να επηρεάζει την πιθανότητα καπνίσματος.

b) Για να είναι ακριβής ο υπολογισμός του διαστήματος εμπιστοσύνης θα πρέπει:

- Τα δείγματα να έχουν ληφθεί με ανεξάρτητο τρόπο από τους 2 πληθυσμούς 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** μεταξύ φύλου και καπνίσματος

Υποθέσεις

1. **Μηδενική Υπόθεση (H0):**
   Το φύλο και το κάπνισμα είναι ανεξάρτητα (δεν υπάρχει σχέση μεταξύ τους).
2. **Εναλλακτική Υπόθεση (Ha):**
   Το φύλο και το κάπνισμα **δεν** είναι ανεξάρτητα (υπάρχει σχέση μεταξύ τους).

Πίνακας Συνάφειας:

```
> addmargins(table_data)
      sex
smoker  F  M Sum
   N   16 18  34
   Y   14 12  26
   Sum 30 30  60
```

d)

```
> chi_test <- chisq.test(table_data, correct=FALSE)
> chi_test

	Pearson's Chi-squared test

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

Το **p-value** που υπολογίστηκε με τον **χ^2-έλεγχο** είναι **0.6023**, το οποίο είναι ίσο με το **p-value** που προέκυψε στον **z-έλεγχο** (*στο ερώτημα (α)*), όπως ήταν αναμενόμενο, δεδομένου ότι ο πίνακας συνάφειας είναι διαστάσεων **2x2**, όπου οι δύο έλεγχοι είναι ισοδύναμοι. Επιπλέον, παρατηρούμε ότι το **χ^2-στατιστικό** (***0.27149***) είναι ίσο με το τετράγωνο του **z-στατιστικού** ($z^2 = 0.52105^2 = 0.27149$), επαληθεύοντας τη μαθηματική σχέση μεταξύ τους. Το αποτέλεσμα αυτό επιβεβαιώνει ότι και οι δύο έλεγχοι οδηγούν στο ίδιο συμπέρασμα, δηλαδή **δεν** απορρίπτεται η μηδενική υπόθεση και **δεν** υπάρχουν στατιστικά σημαντικές ενδείξεις ότι υπάρχει σχέση μεταξύ φύλου και καπνίσματος.

Άσκηση 4

a) Για να είναι ακριβής ο υπολογισμός θα πρέπει:

- Το δείγμα να έχει ληφθεί με SRS από τον πληθυσμό (*πληθυσμός 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_0 : pred ≤ 0.5 "Δεν παρασκευάζονται περισσότερα κόκκινα από μπλε"

H_a: pred > 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 - pnorm(z) = **0.2463583**

Το **p-value** είναι περίπου **25%**, αρκετά μεγαλύτερο από το επίπεδο σημαντικότητας (***α=5%***) γεγονός που σημαίνει ότι **δεν** απορρίπτουμε τη μηδενική υπόθεση ($p_{red} \leq 0.5$). Με βάση αυτό, τα δεδομένα δεν υποστηρίζουν την εναλλακτική υπόθεση ότι παρασκευάζονται περισσότερα κόκκινα smarties από μπλε. Συνεπώς, **δεν** μπορούμε να συμπεράνουμε ότι τα κόκκινα smarties υπερτερούν αριθμητικά των 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)
> p <- c(0.198, 0.178, 0.176, 0.196, 0.252)
>
> chisq.test(d, p = p)

        Chi-squared test for given probabilities

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

Το **p-value** είναι περίπου **2%**, το οποίο είναι **μικρότερο** από το επίπεδο σημαντικότητας (***α=5%***) Αυτό σημαίνει ότι **απορρίπτουμε** τη μηδενική υπόθεση. Συνεπώς, τα δεδομένα παρέχουν στατιστικά σημαντικές ενδείξεις ότι η κατανομή των χρωμάτων έχει **αλλάξει** σε σχέση με το **2009**.

c) **Έλεγχος ομοιογένειας**

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

Ha: 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
```

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**.