3η Άσκηση Στατιστική

ΓΕΩΡΓΙΟΥ ΑΛΕΞΙΟΣ ΛΑΖΑΡΟΣ 3180027

ΒΑΣΙΛΕΙΟΣ-ΕΚΤΩΡ ΚΩΤΣΗΣ-ΠΑΝΑΚΑΚΗΣ 3180094

1)

a.

```
> p = 0.5
> phat = 29/50
> phat
[1] 0.58
> zstar = 1.96 # για 95%

> phat + c(-1,1) * zstar * sqrt(phat*(1-phat)/50)
[1] 0.4431926 0.7168074
```

Διάστημα εμπιστοσύνης εμφάνισης κορώνας 95% = [0.4431926, 0.7168074]

b.

```
> #H0: p = 0.5 H1: p != 0.5
> z <- (phat - p)/sqrt(p*(1-p)/50)
> z
[1] 1.131371
> pvalue <- 2* pnorm(-z)
> pvalue
[1] 0.257899
```

P-value ≈ 0.26 > a = 0.05, άρα η μηδενική υπόθεση δεν απορρίπτεται οπότε με βάση το δείγμα για 95% διάστημα εμπιστοσύνης το νόμισμα είναι δίκαιο.

c.

```
> m <- 0.01
> zstar *zstar *p *(1-p)/(m*m)
[1] 9604
```

Άρα n = 9604

2)

$$\mathrm{C} = 0.95, \mathrm{m} = 0.03$$

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

Δεν γνωρίζουμε τα p στον τύπο αλλά ξέρουμε ότι $p(1-p) \leq \frac{1}{4} \forall\, p \in [0,1]$

Άρα ισχύει ο εξής τύπος για το n

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

Για τα ίδια m,c και z* = 1.96 (λόγω του 95% διαστήματος εμπιστοσύνης)

Έχουμε n ≈ 1067 ≈ 1100, οπότε ουσιαστικά δεν αλλάζει ο αριθμός του δείγματος σε σχέση με την Ελλάδα.

3)

a. Παίρνουμε την μηδενική υπόθεση H_0: $p_1 = p_2$ και την εναλλακτική υπόθεση H_1: $p_1 \neq p_2$, όπου p_1 το ποσοστό των αντρών που καπνίζουν και p_2 το ποσοστό των γυναικών που καπνίζουν του πληθυσμού.

```
> mSmoker
 [1] "YES" "NO"  "NO"  "YES" "YES" "YES" "NO"  "YES" "YES" "NO"  "NO"  "NO"
"YES" "NO"  "NO"  "NO"  "YES" "NO"  "NO"  "NO"  "NO"
[22] "NO"  "YES" "NO"  "YES" "NO"  "YES" "NO"  "YES" "NO"
> wSmoker
 [1] "YES" "YES" "NO"  "NO"  "NO"  "NO"  "NO"  "NO"  "YES" "NO"  "NO"  "NO"
"NO"  "YES" "YES" "YES" "NO"  "YES" "NO"  "NO"  "YES"
[22] "NO"  "YES" "YES" "NO"  "NO"  "YES" "YES" "YES" "YES"
> pm <- length(which(mSmoker=="YES")) / length(mSmoker)
> pm
[1] 0.4
> pw <- length(which(wSmoker=="YES")) / length(wSmoker)
> pw
[1] 0.4666667
> nm <- length(mSmoker)
> nm
[1] 30
> nw <- length(wSmoker)
> nw
[1] 30
> z <- (pm - pw)/sqrt((pm*(1-pm)/nm) +(pw*(1-pw)/nw))
> z
[1] -0.522233
> pvalue <- 2*pnorm(z)
> pvalue
[1] 0.6015081
```

Z = -0.522233

Pvalue = 0.6015081

Δεν απορρίπτεται η μηδενική υπόθεση, το pvalue είναι αρκετά μεγάλο για επίπεδο σημαντικότητας 5%. Άρα δεν υπάρχει κάποια σχέση μεταξύ του φύλου και του αν καπνίζει ένας άνθρωπος.

b.

```
> zstar = 1.96 # για 95%
> (pm - pw) + c(-1,1) * zstar * sqrt(pm*(1-pm)/nm + (pw*(1-pw)/nw))
[1] -0.3168743  0.18
```

Διάστημα εμπιστοσύνης εμφάνισης κορώνας 95% = [-0.3168743, 0.1835410]

c.

```
> chitable <- table(sex,smoker)
> chitable
```

```
       smoker
sex     NO YES
  MAN   18  12
  WOMAN 16  14
> chisq.test(chitable, correct = FALSE)

    Pearson's Chi-squared test

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

Το p-value είναι αρκετά μεγάλο για επίπεδο σημαντικότητας 5%, άρα δεν απορρίπτουμε την μηδενική υπόθεση, άρα οι μεταβλητές sex και smoker είναι ανεξάρτητες.

d.

pvalue (c) = 0.6023 ≈ pvalue(a) = 0.6015

Τα δυο pvalue των υποερωτημάτων a και c είναι περίπου ίσα.

4)

a.

H_0: παρασκευάζονται ίσα μπλε και κόκκινα smarties

H_1: παρασκευάζονται περισσότερα κόκκινα απ' ότι μπλε smarties

```
> y <- c(19,15,80-19-15)
> y
[1] 19 15 46
> nSmartiesPack
[1] 80
> phat1 <- phat2 <- sum(y[1:2])/(2*nSmartiesPack)
> phat3 <- 1 - phat1 - phat2
> phat <- c(phat1,phat2,phat3)

> obs <- sum((y- nSmartiesPack*phat)^2/(nSmartiesPack*phat))
> obs
[1] 0.4705882

> pchisq(obs, df = 2,lower.tail = FALSE)
[1] 0.7903384
```

Υπολογίζουμε τα O_i του δείγματος και τα phat των 3 μεταβλητών της κατηγορικής μεταβλητής. Στην συνέχεια υπολογίζουμε το

$$X^2 \text{ στατιστικό ελέγχου καλής προσαρμογής} = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Και υπολογίζουμε το pvalue με βάση του στατιστικού χ τετράγωνο με βαθμό ελευθερίας ίσο με 2 αφού έχουμε 3 μεταβλητές (κόκκινο, μπλε και άλλο χρώμα).

Pvalue = 0.79

Για 95% διάστημα εμπιστοσύνης, το pvalue είναι αρκετά υψηλό, δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση.

b.

```
> t
 1  2  3  4  5
 8 15 16 19 22

> prop.table(t)

     1      2      3      4      5
0.1000 0.1875 0.2000 0.2375 0.2750

> chisq.test(t, p = c(0.252,0.196,0.176,0.178,0.198), simulate.p.value =
TRUE)
```

```
    Chi-squared test for given probabilities with simulated p-value (based on
2000 replicates)

data:  t
X-squared = 11.613, df = NA, p-value = 0.01849
```

H_0: η σημερινή κατανομή smarties είναι ίδια με του 2009

H_1: η σημερινή κατανομή smarties διαφέρει με την κατανομή του 2009

Έχουμε τον πίνακα συχνότητας των χρωμάτων από το δείγμα και εφαρμόζουμε x τετράγωνο έλεγχο με τις αντίστοιχες πιθανότητες κατανομής του 2009.

Για 95% διάστημα εμπιστοσύνης ,η κατανομή από τότε έχει αλλάξει, το p-value είναι πολύ μικρό για τον έλεγχο που κάναμε. Η μηδενική υπόθεση απορρίπτεται.

c.

```
> Nmnm <- 10+12+20+9+5
> Brownmnm <- 10/Nmnm
> Redmnm <- 12/Nmnm
> Yellowmnm <- 20/Nmnm
> Bluemnm <- 9/Nmnm
> Greenmnm <- 5/Nmnm

> Nmnm
[1] 56

> c(Greenmnm,Bluemnm,Yellowmnm,Redmnm,Brownmnm)
[1] 0.08928571 0.16071429 0.35714286 0.21428571 0.17857143

> chisq.test(t, p = c(Greenmnm,Bluemnm,Yellowmnm,Redmnm,Brownmnm),
simulate.p.value = TRUE)

    Chi-squared test for given probabilities with simulated p-value (based on
2000 replicates)

data:  t
X-squared = 10.358, df = NA, p-value = 0.03348
```

H_0: η κατανομή χρωμάτων smarties και MNMs είναι ίδια

H_1: η κατανομή χρωμάτων smarties και MNMs είναι διαφορετική

Υπολογίζουμε τις πιθανότητες χρωμάτων του δείγματος των MNMs και τις συγκρίνουμε με τις αντίστοιχες πιθανότητες χρωμάτων του δείγματος smarties.

Pvalue = 0.03 άρα για 95% διάστημα εμπιστοσύνης η μηδενική υπόθεση απορρίπτεται, οι κατανομές χρωμάτων είναι διαφορετικές.