# ΣΤΑΤΙΣΤΙΚΗ 3η ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ

ΑΠΟΣΤΟΛΟΠΟΥΛΟΣ ΘΕΜΙΣΤΟΚΛΗΣ p3180013

ΠΑΝΤΕΛΟΠΟΥΛΟΣ ΣΤΥΛΙΑΝΟΣ p3180142

1)

A)

Καθώς X=29>15 και n-X=31>15 και τα δεδομένα μας έρχονται από ένα απλό τυχαίο δείγμα θα χρησιμοποιήσουμε για τον υπολογισμό(για «ακριβή» διάστημα εμπιστοσύνης)τον τύπο:

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

(με z*=1.96 καθώς ζητείται 95% διάστημα εμπιστοσύνης)

Από την R έχουμε:

```
> z<-1.96
> pstar<-29/50
> pstar
[1] 0.58
> sdp<-sgrt((pstar*(1-pstar))/50)
Error in sgrt((pstar * (1 - pstar))/50) : could not find functio
> sdp<-sqrt((pstar*(1-pstar))/50)
> sdp
[1] 0.06979971
> x<-pstar+c(-1,1)*z*sdp
> x
[1] 0.4431926 0.7168074
>
```

Άρα έχουμε το διάστημα 0.4431926,0.7168074

B)Παίρνουμε ως h0:p=0.5 και ha:p!=0.5

Θα κάνουμε έλεγχο σημαντικότητας χρησιμοποιώντας στατιστικό έλεγχο Z (Z-Έλεγχο σημαντικότητας).Ο ακόλουθος έλεγχος θα είναι αρκετά «ακριβής» καθώς οι σχέσεις n*p0>10 και n(1-p0)>10 ισχύουν για n=50, p0=0.5(και τα δεδομένα μας είναι από απλό τυχαίο δείγμα)

.Κάνοντας τον έλεγχο στην R έχουμε τα ακόλουθα αποτελέσματα:

```
> z_score<-(pstar-0.5)/sqrt((0.5*(1-0.5))/50)
> s_score
Error: object 's_score' not found
> z_score
[1] 1.131371
> p_val<-2*pnorm(-abs(z_score))
> p_val
[1] 0.257899
```

Για επίπεδο σημαντικότητας 5% έχουμε πως pvalue>a=5% άρα η μηδενική μας υπόθεση(ότι το νόμισμα είναι δίκαιο) στέκει.

C)

Θα χρησιμοποιήσουμε τον τύπο:

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

Εφόσον το p=1/2 ο παραπάνω τύπος γίνεται:

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

Από R έχουμε:

```
> z^2/(4*0.01^2)->y
> y
[1] 9604
> n>=9604
```

(για Z=1.96 καθώς μιλάμε για διάστημα εμπιστοσύνης 95% και m=0.01) άρα θέλουμε n>=9604 δηλαδή τουλάχιστον 9604 ρίψεις.

2)

Εφόσον δεν έχουμε κάποια αλλαγή στο διάστημα εμπιστοσύνης και στο περιθώριο σφάλματος ο αριθμός των ατόμων που θα πάρουν μέρος στην δημοσκόπηση για την Αμερική θα είναι ο ίδιος όπως και στην Ελλάδα. Παρόλο την μεγάλη διαφορά στον αριθμό του πληθυσμού των 2 χωρών εμείς απλά ψάχνουμε πόσα άτομα(παράμετρος n) θα πρέπει να έχουμε στο «πείραμα» μας(εδώ πείραμα = δημοσκόπηση»). Αυτό δεν επηρεάζεται από τον συνολικό μας πληθυσμό αλλά από το διάστημα εμπιστοσύνης και το περιθώριο σφάλματος. Φανταστείτε ότι είστε σε ένα εργαστήριο και θέλετε να κάνετε κάποιο πείραμα σε πληθυσμό ποντικιών για να δείτε κάποια επίδραση. Έστω ότι υπάρχουν 2 πληθυσμοί ποντικιών, ο ένας πολύ μεγαλύτερος σε αριθμό από τον άλλον. Αν οι μεταβλητές διαστήματος εμπιστοσύνης και περιθώριου λάθους παραμείνουν αμετάβλητες και για τους 2 πληθυσμούς, για να έχω τα αποτελέσματα που θέλω θα πρέπει να πάρω τον ίδιο αριθμό ποντικιών και να τους κάνω το πείραμα μου ΚΑΙ ΣΤΟΥΣ 2 πληθυσμούς.

3)

A)

Βάζοντας τα δεδομένα στην R και με την χρήση του τύπου

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \quad \text{όπου } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Παίρνουμε από την R:

```
> attach(x)
> p1cap<-with(x[sex=="MALE",], mean(smoker=="YES"))
> p1cap
[1] 0.4
> p2cap<-with(x[sex=="FEMALE",], mean(smoker=="YES"))
> p2cap
[1] 0.4666667
> pcap<-mean(smoker == "YES")
> z<-(p2cap-p1cap)/sqrt(pcap*(1-pcap)*(1/sum(sex=="MALE") + 1/sum(sex=="FEMALE"$
> z
[1] 0.5210501
> #pososto antrwn kapnistwn = gynaikwn p1=p2 diaforetika p1!=p2
> 2*pnorm(-abs(z))
[1] 0.6023319
> #ortho to p1=p2
> 
```

Με h0:p1=p2 ha:p1!=p2 με p1 ποσοστό αντρών καπνιστών και p2 ποσοστό γυναικών καπνιστών.

Έχουμε αρκετά υψηλό pvalue(0.6023319) επομένως δεν θα απορρίψουμε την υπόθεση μας.Το pvalue είναι ακριβές καθώς έχουμε δεδομένα από τυχαίο δείγμα το οποίο είναι και αρκετά μεγάλο

Β)

Με χρήση του τύπου:

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

Για z*=1.96(95% διάστημα εμπιστοσύνης) έχουμε από την R:

```
> #3β
> p2cap - p1cap + c(-1,1)*1.96*sqrt(p2cap*(1-p2cap)/sum(sex=="FEMALE")+p1cap*(1-p1cap)/sum(sex=="MALE"))
[1] -0.1835410  0.3168743
> 
```

Επομένως έχουμε το διάστημα -0.1835410, 0.3168743

C KAI D)

Θα πραγματοποιήσουμε τον έλεγχο στην R.Έχουμε ότι:

```
> table(sex, smoker)
        smoker
sex      NO YES
  FEMALE 16  14
  MALE   18  12
> t<-table(sex,smoker)
> chisq.test(t)

        Pearson's Chi-squared test with Yates' continuity correction

data:  t
X-squared = 0.067873, df = 1, p-value = 0.7945

> chisq.test(t,correct = FALSE)

        Pearson's Chi-squared test

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

> #auta gia 3c 3d blepw oti p value me x tetragwno test = p value z elegxoy
```

(πάνω πάνω έχουμε τον πίνακα συνάφειας, ο πρώτος έλεγχος γίνεται με δίορθωση του Yates , ενώ ο δεύτερος έλεγχος δεν γίνεται με κάποια διόρθωση.)

Με τον έλεγχο αυτό εξετάζουμε αν οι 2 πλυθυσμοί μας(άντρες,γυναίκες καπνιστές) έχουν «παρόμοια» κατανομή.Βλέπουμε ότι έχουμε p value μεγάλο(και στις 2 περιπτώσεις) άρα έχουμε πάτημα στην υπόθεση μας ότι έχουμε παρόμοια κατανομή στους πληθυσμούς μας.

Στον δεύτερο έλεγχο(χωρίς την διόρθωση) το pvalue είναι ίδιο με το pvalue στον έλεγχο Z στο ερώτημα B.

4)

A)

Πραγματοποιούμε Z έλεγχο με τον τύπο:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \quad \text{όπου } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Από την R έχουμε:

```
> pcap<-mean(Color == "red" | Color == "blue")
> predcap<-mean(Color == "red")
> predcap
[1] 0.2375
> pbluecap<-mean(Color == "blue")
> pbluecap
[1] 0.1875
> z<(predcap-pbluecap)/ sqrt(pcap*(1-pcap)* (1/sum(Color == "red") + 1/sum(Color == "blue")))
[1] TRUE
> z<-(predcap-pbluecap)/ sqrt(pcap*(1-pcap)* (1/sum(Color == "red") + 1/sum(Color == "blue")))
> z
[1] 0.2928361
> 1-pnorm(abs(-z))
[1] 0.3848237
```

Με h0:pred<=pblue,ha:pred>pblue(ha αυτό που θέλουμε να δούμε)

Με p value αρκετά μεγάλο(38%) δεν θα απορρίψουμε την μηδενική μας υπόθεση άρα δεν ισχύει ότι κατασκευάζονται περισσότερα κόκκινα από μπλέ σμαρτις.

B)Θα πραγματοποιήσουμε X^2 έλεγχο στα δεδομένα μας αλλά με παράμετρο ποσοστού για το κάθε «σμαρτι» όχι ισότιμη με τις άλλες(20%) αλλά με τα ποσοστα που μας δίνονται στην άσκηση.

Από την R έχουμε:

```
> ask4b
x
  blue  brown  green    red yellow
    15     22      8     19     16
> chisq.test(ask4b, p=c(0.196, 0.198, 0.252, 0.178, 0.176))

        Chi-squared test for given probabilities

data:  ask4b
X-squared = 11.613, df = 4, p-value = 0.02048

> #p=c(ποσοστο μπλε, ποσοστο καφε, ποσοστο πρασινων, ποσοστο κοκκινων, ποσοστο κιτρινων
> |
```

Παρατηρούμε ότι έχουμε p value = 2%.Η υπόθεση ότι τα δείγματα μας είναι κατανεμημένα έτσι ώστε το καθένα να έχει ένα τέτοιο ποσοστό(19.8% τα καφέ πχ) είναι στην κρίση μας αν θα απορριφθεί η όχι(χωρίς να έχουμε κάποιο επίπεδο σημαντικότητας α).Εμείς θεωρούμε το pvalue αρκετά μικρό για να απορρίψουμε την υπόθεση μας.Άρα η κατανομή δεν έχει παραμείνει ίδια

C)

Θα χρησιμοποιήσουμε παρόμοιο σκεπτικό με το ερώτημα B).

Αρχικά θα υπολογίσουμε το ποσοστό κάθε χρώματος σμάρτι από το πακέτο m&ms.(xi/56 όπου xi = #σμαρτι στο πακέτο m&ms i χρώματος για i=καφέ,κόκκινο,κίτρινο,μπλέ,πράσινο)

Τα ποσοστά αυτά βγαίνουν ως εξής(με στρογγυλοποιήσεις):

0.18, 0.21, 0.36, 0.16, 0.09 και το sum τους = 1

.Άρα θα κάνουμε χ τετράγωνο έλεγχο στα δεδομένα μας με τα παραπάνω ποσοστά και με μηδενική υπόθεση ότι η κατανομή τους είναι παρόμοια(δηλαδή παρόμοια με m&ms).Απο R έχουμε:

```
> ask4b
x
  blue  brown  green    red yellow
    15     22      8     19     16
> chisq.test(ask4b, p=c(9/56, 10/56, 5/56, 12/56, 20/56))

        Chi-squared test for given probabilities

data:  ask4b
X-squared = 10.358, df = 4, p-value = 0.03481

> 
```

Βλέπουμε πολύ μικρό p value, άρα η μηδενική μας υπόθεση απορρίπτεται.