ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΦΟΡΙΚΗ

ΤΣΙΝΤΖΟΣ ΙΩΑΝΝΗΣ p3200211

ΜΗΤΣΑΝΑΣ ΚΩΝΣΤΑΝΤΙΝΟΣ p3200103

3Η ΕΡΓΑΣΙΑ

ΑΣΚΗΣΗ 1

1. Για ένα νόμισμα θέλετε να διαπιστώσετε εάν είναι «δίκαιο», δηλαδή εάν η συχνότητες εμφάνισης κορώνας και γραμμάτων -εάν πραγματοποιούσατε άπειρες ρίψεις- θα ήταν ίσες. Πραγματοποιείτε n = 50 ρίψεις και εμφανίζονται 29 κορώνες.

a. Δώστε ένα 95% διάστημα εμπιστοσύνης για τη συχνότητα εμφάνισης κορώνας.

b. Τι συμπεράνετε για το εάν το νόμισμα είναι δίκαιο σε επίπεδο σημαντικότητας 5%;

c. Πόσες ρίψεις θα έπρεπε να πραγματοποιήσετε εάν το περιθώριο λάθους στο διάστημα του ερωτήματος (a) θα θέλατε να είναι μικρότερο του 1%;

A)

Πραγματοποιούμε άπειρες ρίψεις και παίρνουμε δείγμα από αυτές με n = 50 ρίψεις. Έστω X οι κορώνες και Y τα γράμματα με X = 29 και Y = 21. X > 15 και Y >15 άρα τα δεδομένα μας είναι κατάλληλα.

Το 95% διάστημα εμπιστοσύνης είναι $\hat{p} \pm z*\surd\ (\hat{p}(1-\hat{p})\ /n)$ = [0.443,0.717] όπου χρησιμοποιήσαμε $z*$ = 1.96 για επίπεδο εμπιστοσύνης 95%.

```
> n <- 50
> x <- 29
> p_hat <- x/n
> z <- qnorm(0.975)
> p_hat + c(-1,1)*z*sqrt(p_hat*(1-p_hat)/n)
[1] 0.4431951 0.7168049
```

Β)

Δίπλευρος έλεγχος H0 : p = 0.5

Αφού το ζάρι είναι δίκαιο άρα η συχνότητα εμφάνισης κορώνας είναι 1/2. Στατικός έλεγχος $z = (0.58-0.5) / \sqrt{(0.5(1-0.5) / 50)} = 1.1313$ άρα p-value =

$2\Phi(-|z|) = 0.25$.

Το p-value είναι μικρή τιμή άρα δε θα μπορούσαμε να απορρίψουμε την μηδενική υπόθεση με σιγουριά.

```
n <- 50
x <- 29
p_hat <- x/n
z <- (p_hat - 0.5) / sqrt(0.5*(1-0.5)/n)
p_value <- 2*pnorm(-abs(z))
if (p_value < 0.05) {
    print("Reject the null hypothesis")
} else {
    print("Fail to reject the null hypothesis")
}
1] "Fail to reject the null hypothesis"
```

Γ)

Αρκούν 9604 ρίψεις.

```
> E <- 0.01
> p <- 0.5
> z <- qnorm(1-0.05/2)
> n <- (z^2 * p * (1-p)) / (E^2)
> ceiling(n)
[1] 9604
```

ΑΣΚΗΣΗ 2

Στην Ελλάδα (με πληθυσμό 10 εκατομμυρίων περίπου) οι δημοσκοπήσεις εκτίμησης ποσοστού ψηφοφόρων χρησιμοποιούν γύρω στα 1100 άτομα για την κατασκευή 95% διαστημάτων εμπιστοσύνης με περιθώριο σφάλματος 3%. Πόσα άτομα απαιτούνται για την πραγματοποίηση αντίστοιχων δημοσκοπήσεων στις Η.Π.Α (με πληθυσμό περίπου 300 εκατομμυρίων) για κατασκευή διαστημάτων εμπιστοσύνης με το ίδιο περιθώριο σφάλματος και επίπεδο εμπιστοσύνης; Δικαιολογήστε την απάντησή σας.

Το μέγεθος του δείγματος που χρησιμοποιείται για τις δημοσκοπήσεις είναι ανεξάρτητο από το μέγεθος του πληθυσμού, καθώς σύμφωνα με τον τύπο

$n \geq (z* \ \hat{}2) \ / \ (4m\hat{}2)$, το πρώτο εξαρτάται μόνο από το περιθώριο λάθους m και το επίπεδο εμπιστοσύνης.

Άρα και για τις δημοσκοπήσεις στις Η.Π.Α ο αριθμός των 1100 ατόμων είναι αρκετός.

ΑΣΚΗΣΗ 3

Εδώ θα εξετάσουμε τη σχέση μεταξύ φύλου και καπνίσματος σε έναν πληθυσμό από όπου λαμβάνεται ένα απλό τυχαίο δείγμα που δίδεται στον Πίνακα 1.

a. Διατυπώστε έναν z έλεγχο υπόθεσης που να εξετάζει την ύπαρξη σχέσης μεταξύ φύλου και καπνίσματος και εφαρμόστε τον στα δεδομένα του Πίνακα 1.

b. Δώστε ένα 95% διάστημα εμπιστοσύνης για τη διαφορά του ποσοστού καπνιστών μεταξύ ανδρών και γυναικών.

c. Διατυπώστε έναν χ 2 έλεγχο σημαντικότητας που να εξετάζει την ύπαρξη σχέσης μεταξύ φύλου και καπνίσματος και δώστε ένα πίνακα συνάφειας βάσει των δεδομένων του Πίνακα 1.

d. Δώστε το p value του χ 2 ελέγχου και συγκρίνετε με το ερώτημα (a).

A)

Για να ελέγξουμε την ύπαρξη σχέσης μεταξύ φύλου και καπνίσματος, θα θεωρήσουμε τον δίπλευρο έλεγχο H0: p1 = p2, όπου p1 και p2 είναι το ποσοστό των καπνιστών στον υποπληθυσμό των ανδρών και γυναικών αντίστοιχα.

Έχουμε n1 = 30 το πλήθος των αντρών και n2 = 30 το πλήθος των γυναικών.

Από τα δεδομένα του Πίνακα 1, υπολογίζουμε ότι p1 = 0.4 και p2 = 0.4667. Επίσης, θεωρούμε ότι n1 = 30, X1 = 12, n2 = 30, X2 = 14. Χρησιμοποιώντας τους όρους αυτούς και υπολογίζοντας το z-value, θα έχουμε

z = (0.4 - 0.4667) / sqrt(0.4333 * (1 - 0.4333) * (1/30 + 1/30)) = -0.521.

Στη συνέχεια , μπορούμε να βρούμε το p-value που συνδέεται με αυτό το z-value.

Συνεπώς, θα έχουμε p-value ≈ 0.6, οπότε δεν υπάρχει σημαντική σχέση μεταξύ φύλου και καπνίσματος

```
> # Calculate sample sizes and proportions
> n1 <- sum(data$GENDER == "A")
> x1 <- sum(data$GENDER == "A" & data$SMOKER == "NAI")
> p1 <- x1/n1
> n2 <- sum(data$GENDER == "Γ")
> x2 <- sum(data$GENDER == "Γ" & data$SMOKER == "NAI")
> p2 <- x2/n2
>
> # Calculate pooled proportion and standard error
> p <- (x1 + x2)/(n1 + n2)
> se <- sqrt(p * (1 - p) * (1/n1 + 1/n2))
>
> # Calculate Z-score and P-value
> z <- (p1 - p2)/se
> p_value <- 2 * (1 - pnorm(abs(z)))
>
> # Print results
> print(paste0("Z-score: ", round(z, 3)))
[1] "Z-score: -0.521"
> print(paste0("P-value: ", round(p_value, 6)))
[1] "P-value: 0.602332"
```

B)

Παρατηρούμε ότι $X1 \geq 10$, $n1 - X1 \geq 10$,$X2 \geq 10$, $n2 - X2 \geq 10$,

Αρα το επίπεδο εμπιστοσύνης του διαστήματος που δίδεται παρακάτω είναι ακριβές

(95%): $\hat{p}1 - \hat{p}2 \pm z* \surd ((\hat{p}1\ (1 - \hat{p}1\ )/\ n1) + (\hat{p}2\ (1 - \hat{p}2\ )\ /n2)\ )$= [−0.00015, -0.133]

```
> low<- p1-p2-z* sqrt((p1 * (1 - p1)/n1)  +  (p2 * (1 - p2)/n2))
> low
[1] -0.0001510006
> high<- p1-p2+z* sqrt((p1 * (1 - p1)/n1)  +  (p2 * (1 - p2)/n2))
> high
[1] -0.1331823
```

C)

Για να ελέγξουμε την ύπαρξη σχέσης μεταξύ φύλου και καπνίσματος με χρήση ενός χ^2 ελέγχου σημαντικότητας, θα πρέπει να δημιουργήσουμε έναν πίνακα συνάφειας βάσει των δεδομένων του Πίνακα 1.

Έστω ο έλεγχος

H0 : Το φύλο δεν έχει σχέση με το κάπνισμα

Ha : Το φύλο έχει σχέση με το κάπνισμα

ΠΙΝΑΚΑΣ ΣΥΝΑΦΕΙΑΣ

ΦΥΛΟ	ΚΑΠΝΙΣΤΗΣ	ΜΗ ΚΑΠΝΙΣΤΗΣ	
ΑΝΤΡΕΣ	12	18	30
ΓΥΝΑΙΚΕΣ	14	16	30
	26	34	60

D)

ΦΥΛΟ	ΚΑΠΝΙΣΤΗΣ	ΜΗ ΚΑΠΝΙΣΤΗΣ	
ΑΝΤΡΕΣ	13	17	30
ΓΥΝΑΙΚΕΣ	13	17	30
	26	34	60

Στατιστικός έλεγχος $X^2 = (12-13)^2/13 + (18-17)^2 / 17 + (14-13)^2 / 13 + (16-17)^2 / 17 \approx 0.07 + 0.058 + 0.07 + 0.058 \approx 0.256$

p value=0.6

ΑΣΚΗΣΗ 4

Παρασκευάζονται περισσότερα κόκκινα smarties (χρωματιστά σοκολατένια κουφετάκια) από ότι μπλε; Αγοράζετε μια συσκευασία από το περίπτερο όπου βρίσκετε 22 καφέ, 19 κόκκινα, 16 κίτρινα, 15 μπλε και 8 πράσινα κουφέτα.

a. **Απαντήστε στο παραπάνω ερώτημα εφαρμόζοντας ένα έλεγχο σημαντικότητας.**
b. **Το 2009 είχε μετρηθεί με μεγάλη ακρίβεια το ποσοστό εμφάνισης των χρωμάτων καφέ, κόκκινο, κίτρινο, μπλε και πράσινο, το οποίο βρέθηκε ότι ήταν 19.8%, 17.8%, 17.6%, 19.6%, 25.2% αντίστοιχα. Έχει αλλάξει η κατανομή αυτή από τότε;**
c. **Η αναλογία χρωμάτων στα smarties είναι ίδια με αυτή στα M&Ms (άλλο προϊόν χρωματιστών σοκολατένιων κουφέτων); Ανοίγοντας μια συσκευασία M&Ms βρίσκετε 10 καφέ, 12 κόκκινα, 20 κίτρινα, 9 μπλε και 5 πράσινα.**

A)

Θέλουμε να μάθουμε αν παρασκευάζονται περισσότερα κόκκινα smarties από ότι μπλε, άρα μας ενδιαφέρει ο υποπληθυσμός των κόκκινων και μπλε smarties.

Έστω p το ποσοστό των κόκκινων και 1-p των μπλε κουφέτων. Ενδιαφερόμαστε να μάθουμε αν τα κόκκινα smarties είναι περισσότερα οπότε θα πάρουμε μονόπλευρο έλεγχο υπόθεσης

H0 : p = 1/2 ,

Ha : p > 1/2 .

Από τα δεδομένα της εκφώνησης έχουμε ότι:

n = 19 + 15 = 34, X = 19 και

$\hat{p}= X/n$ = 19/34 = 0.5588.

Θεωρούμε ότι η συσκευασία αποτελεί ένα τυχαίο δείγμα με τα smarties που φτιάχνονται, δηλαδή ότι ο τρόπος με τον οποίο αναμιγνύονται τα χρώματα είναι τυχαίος.

Ο στατιστικός έλεγχος είναι z= $\hat{p}$−0.5 √ 0.5(1−0.5) n = 0.686 και το p-value = 2Φ(- |z|)=0.4902.

Αφού η τιμή του p-value δεν είναι πολύ μικρή, η μηδενική υπόθεση είναι αποδεκτή.

Άρα, δεν υπάρχει σημαντική διαφορά στον αριθμό των κόκκινων και μπλε κουφέτων.

```
> H0 <- 0.5
> Ha <- ">"
> n <- 34
> X <- 19
> p_hat <- X/n
> z <- (p_hat - H0) / sqrt(H0 * (1 - H0) / n)
> p_value <- pnorm(z, lower.tail = F)
> alpha <- 0.05
> if (p_value < alpha) {
+ print("Reject the null hypothesis")
+ } else {
+ print("Fail to reject the null hypothesis")
+ }
[1] "Fail to reject the null hypothesis"
>
> if (p_value < alpha) {
+ print("There is evidence to suggest that there are more red smarties than blue smarties.")
+ } else {
+ print("There is not enough evidence to suggest that there are more red smarties than blue smarties.")
+ }
[1] "There is not enough evidence to suggest that there are more red smarties than blue smarties."
```

B)

Θέλουμε να δούμε αν η κατανομή των χρωμάτων είναι η ίδια σε σχέση με αυτή του 2009. Θα εφαρμόσουμε τον χ 2 έλεγχο καλής προσαρμογής:

H0: η κατανομή των χρωμάτων καφέ, κόκκινο, κίτρινο, μπλε και πράσινο είναι 19.8%, 17.8%, 17.6%, 19.6% και 25.2% αντίστοιχα,

Ha: η κατανομή των χρωμάτων είναι διαφορετική. Από τα δεδομένα προκύπτει ο ακόλουθος πίνακας:

	Δεδομένα	Αναμενόμενες Τιμές
Καφέ	22	15.84
Κόκκινο	19	14.24
Κίτρινο	16	14.08
Μπλέ	15	15.68
Πράσινο	8	20.16

Με τη βοήθεια της R βρίσκουμε ότι p-value= 0.02, το οποίο είναι αρκετά μικρό για να ισχύει η μηδενική υπόθεση.

Άρα η κατανομή των χρωμάτων είναι διαφορετική σε σχέση με το 2009.

C)

Θα θεωρήσουμε ότι η συσκευασία των M&Ms αποτελεί ένα τυχαίο δείγμα από τον πληθυσμό των κουφέτων της συγκεκριμένης μάρκας. Θα ελέγξουμε αν τα δείγματα των smarties και των M&Ms προήλθαν από πληθυσμούς με την ίδια κατανομή χρωμάτων.

Για αυτό τον έλεγχο θα χρησιμοποιήσουμε έναν χ^2 έλεγχο για ομοιογένεια.

H0: οι πληθυσμοί των smarties και των M&Ms είναι ομοιογενείς

Hα: οι πληθυσμοί δεν είναι ομοιογενείς.

Από τα δεδομένα προκύπτει ο ακόλουθος πίνακας συνάφειας:

	smarties	M&Ms	
Καφέ	22	10	32
Κόκκινο	19	12	31
Κίτρινο	16	20	36
Μπλέ	15	9	24
Πράσινο	8	5	13
	80	56	136

Με τη βοήθεια της R βρίσκουμε ότι ο στατιστικός έλεγχος είναι χ^2 = 4.626, οι βαθμοί ελευθερίας = 4 και p-value= 0.3278.

Το p-value είναι αρκετά μεγάλο ώστε να μην απορριφθεί η μηδενική υπόθεση, οπότε οι κατανομές των χρωμάτων στα smarties και τα M&Ms είναι ίδιες.