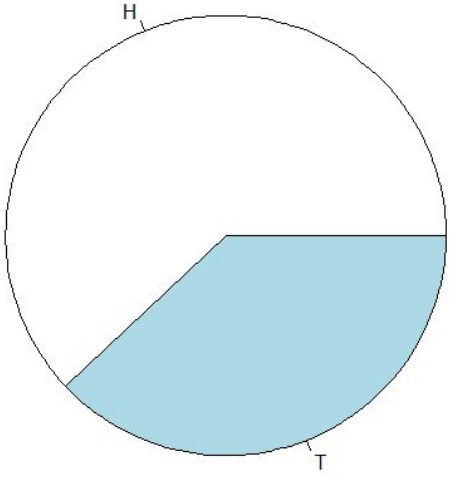
Στατιστική στην Πληροφορική
2020-2021

Φέκα Αγγελική Α.Μ:3140290
Σταυρουλάκη Μαρία Α.Μ:3160168

3η σειρά ασκήσεων

Άσκηση 1η :

a) Πριν εξάγουμε τα συμπεράσματά μας, πρέπει να ελέγξουμε αν τα δεδομένα μας είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε. Προκειμένου να εξάγουμε το συμπέρασμα αυτό, εφαρμόζουμε διερευνητική ανάλυση στα δεδομένα αυτά. Παρακάτω εμφανίζεται το διάγραμμα pie των δεδομένων μας:



Όπου με το γράμμα **H** συμβολίζουμε τις κορώνες*(heads)* και με το γράμμα **T** τα γράμματα*(tails)*.
Από το παραπάνω διάγραμμα συμπεραίνουμε ότι τα δεδομένα μας είναι κατάλληλα για τις μεθόδους συμπερασματολογίας, καθώς το δείγμα που μας δίνεται(n = 50 ρίξεις) είναι ένα τυχαίο δείγμα από το σύνολο των άπειρων ρίψεων. Επιπλέον δεν υπάρχουν ατυπικά σημεία που να επηρεάζουν το αποτέλεσμα, η τυπική απόκλιση είναι γνωστή και τέλος ο δειγματικός μέσος κατανέμεται κανονικά.

Από τα δεδομένα μας έχουμε: x = 31 κορώνες από τις n = 50 ρίψεις.

Επίσης έχουμε ότι το $x \geq 15$ και ισχύει ότι $n - x = 19 \geq 15$, άρα συμπεραίνουμε ότι τα 95% διαστήματα εμπιστοσύνης είναι αρκετά ακριβή.

Το 95% διάστημα εμπιστοσύνης είναι $\widehat{p} \pm z_* \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} = [0.4855,\ 0.7545]$, όπου χρησιμοποιήσαμε $z_* = 1.96$ για επίπεδο εμπιστοσύνης 95%.

b) Θα εφαρμόσουμε τον δίπλευρο έλεγχο H_0 : p = 0.5, όπου p είναι η συχνότητα εμφάνισης κορώνας σε άπειρες ρίψεις.

Το στατιστικό ελέγχου z είναι $z = \frac{\widehat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}} = 1.697056$ με p-value = 0.0897.

Επομένως, συμπεραίνουμε ότι το p-value είναι ναι μεν μικρό, αλλά όχι τόσο ώστε να απορρίψουμε τη μηδενική υπόθεση με σιγουριά.

c) Εάν το περιθώριο λάθους του ερωτήματος a) είναι μικρότερο του 1% και το διάστημα εμπιστοσύνης εξακολουθεί να είναι C = 95%, τότε ισχύει ότι C = $\Phi(z^*) - \Phi(-z^*)$ με z* = 1.96.

Άρα ο αριθμός του τυχαίου δείγματος πρέπει να είναι $n \geq \frac{z_*^2}{4 \times 0.01^2} = 9603.647$

Επομένως θα πρέπει να πραγματοποιήσουμε 9604 ρίψεις προκειμένου το περιθώριο λάθους στο διάστημα του ερωτήματος a) να είναι μικρότερο του 1%.

Άσκηση 2η :

Γνωρίζουμε ότι το μέγεθος του δείγματος εξαρτάται μόνο από το περιθώριο λάθους m και δεν εξαρτάται από το μέγεθος του πληθυσμού. Επιπλέον το επίπεδο εμπιστοσύνης καθορίζεται από το z_*

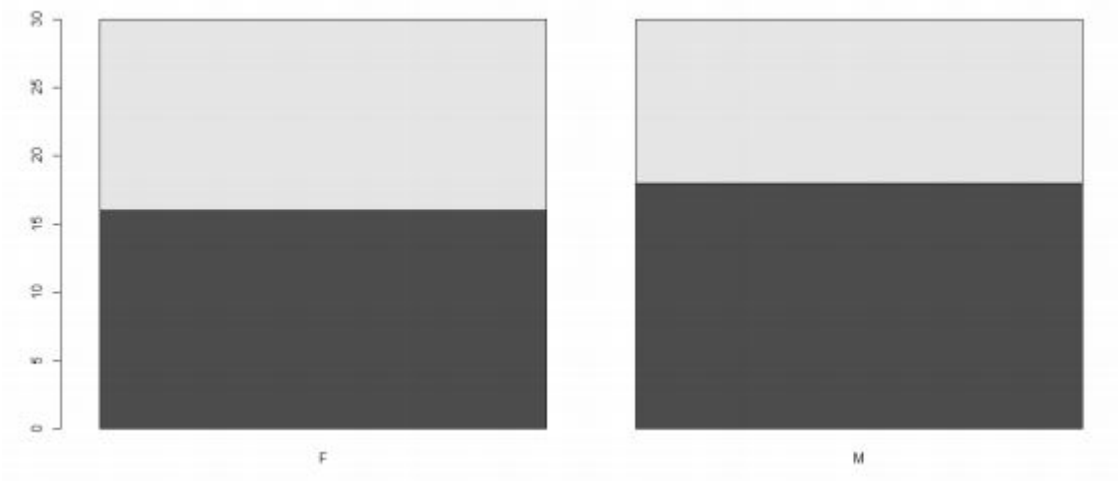
Επομένως χρησιμοποιούμε τον τύπο $n \geq \frac{z_*^2}{4 \times m^2}$ και βρίσκουμε ότι για την πραγματοποίηση των αντίστοιχων δημοσκοπήσεων στις Η.Π.Α χρειάζονται 1100 άτομα αρκούν.

Άσκηση 3η :

a) Μας ζητείται να αποδείξουμε ότι υπάρχει σχέση μεταξύ φύλου και καπνίσματος (μηδενική υπόθεση H_0). Υπολογίζοντας το p-value του z ελέγχου υπόθεσης με χρήση της R προκύπτει: p-value = 0.6023319. Επομένως, εφόσον το p-value είναι μεγάλο δεν απορρίπτουμε τη μηδενική μας υπόθεση,

δηλαδή δεν φαίνεται να υπάρχει σχέση μεταξύ φύλου και καπνίσματος στον πληθυσμό αυτό.

b) Αρχικά κάνουμε διερευνητική ανάλυση στα δεδομένα προκειμένου να δούμε αν είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε. Παρακάτω εμφανίζεται το διάγραμμα barplot των δεδομένων μας:



Όπου F συμβολίζονται οι γυναίκες και M οι άντρες. Με σκούρο χρώμα στο διάγραμμα απεικονίζεται ο αριθμός των καπνιστών και με ανοιχτό ο αριθμός των μη καπνιστών. Μέσω του barplot βλέπουμε ότι τα δεδομένα μας είναι κατάλληλα για τις μεθόδους συμπερασματολογίας, διότι το δείγμα είναι τυχαίο, δηλαδή πρόκειται για τυχαιοποιημένη και όχι αυτοματοποιημένη διαδικασία, δεν υπάρχουν ατυπικά σημεία που επηρεάζουν το αποτέλεσμα, η τυπική απόκλιση είναι γνωστή και ο δειγματικός μέσος κατανέμεται κανονικά.

Με την βοήθεια λογισμικού στατιστικών υπολογισμών βρήκαμε ότι το 95% διάστημα εμπιστοσύνης για τη διαφορά του ποσοστού καπνιστών μεταξύ ανδρών και γυναικών είναι [−0.194, 0.061].

c) Η μηδενική υπόθεση H_0 χαρακτηρίζει όπως και σε προηγούμενο ερώτημα την περίπτωση που το φύλλο δε σχετίζεται με το κάπνισμα, άρα παίρνουμε Ha την υπόθεση ότι σχετίζεται. Ο πίνακας συνάφειας δεδομένων είναι ο εξής:

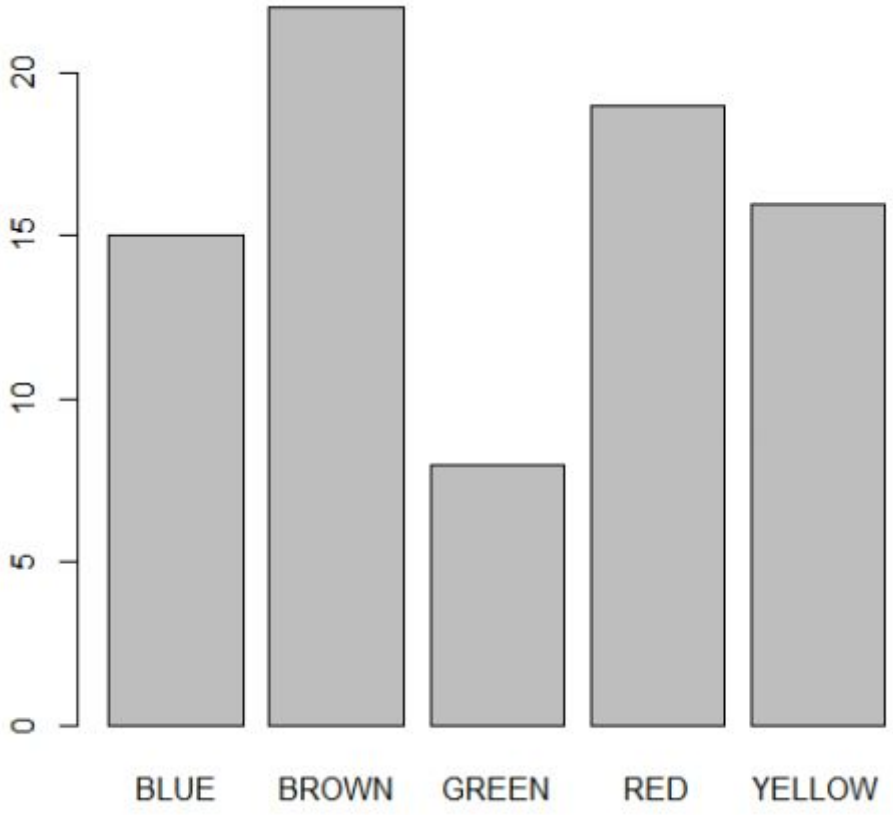
	ΓΥΝΑΙΚΕΣ	**ΑΝΤΡΕΣ**	**ΣΥΝΟΛΟ**
ΚΑΠΝΙΣΤΕΣ	14	12	26
ΜΗ ΚΑΠΝΙΣΤΕΣ	16	18	34
ΣΥΝΟΛΟ	30	30	60

d) Το p value του χ^2 ελέγχου είναι ακριβές, διότι η παραγωγή δεδομένων είναι κατάλληλη, δηλαδη διαθέτουμε απλό τυχαίο δείγμα μεγέθους n και για 2×2 πίνακες: $E_{i,j} \geq 5$ για κάθε i, j. Με τη χρήση του προγράμματος της R προκύπτει ότι το p-value του χ^2 ελέγχου είναι 0.6023.

Παρατηρούμε ότι το p-value είναι το ίδιο με το p-value που βρήκαμε στον z έλεγχο του πρώτου υποερώτηματος(a). Άρα οι δύο έλεγχοι z είναι ισοδύναμοι με χ^2 ελεγχο, σε 2×2 πίνακες συνάφειας.

Άσκηση 4η :

a) Αρχικά πραγματοποιούμε διερευνητική ανάλυση των δεδομένων μας για να διαπιστώσουμε εάν τα δεδομένα αυτά είναι κατάλληλα για τις γνωστές μεθόδους συμπερασματολογίας. Χρησιμοποιώντας το λογισμικό R, παίρνουμε το παρακάτω διάγραμμα barplot μέσω του οποίου βλέπουμε ότι τα δεδομένα μας είναι κατάλληλα για τις μεθόδους συμπερασματολογίας, διότι το δείγμα είναι τυχαίο, δηλαδή πρόκειται για τυχαιοποιημένη και όχι αυτοματοποιημένη διαδικασία και το μέσο πλήθος “επιτυχιών” και “αποτυχιών” κάτω από την H_0 είναι πάνω από 10.

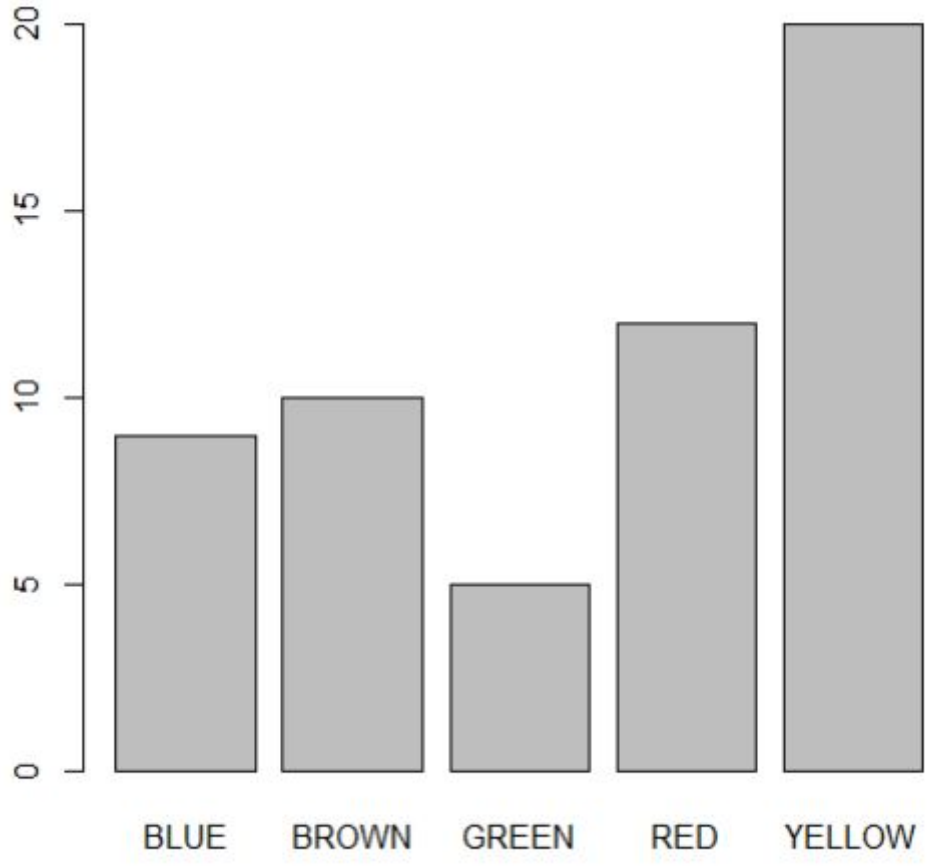


Η αρχική μας υπόθεση H_0 είναι ότι παρασκευάζονται περισσότερα κόκκινα κουφέτα από ότι μπλέ, ενώ η εναλλακτική υπόθεση H_α είναι ότι δεν παρασκευάζονται περισσότερα. Κάνοντας έλεγχο σημαντικότητας βλέπουμε ότι το p-value ισούται με: 0.4427, η οποία δεν είναι μικρή και άρα δεν

απορρίπτεται η μηδενική υπόθεση και η διαφορά ανάμεσα και κόκκινα και μπλε δεν είναι σημαντική.

b) Παρατηρούμε ότι σε σχέση με το 2009 έχει αλλάξει η κατανομή καθώς το ποσοστό των καφέ, των κόκκινων και των κίτρινων κουφέτων μειώθηκε, ενώ των μπλε και πράσινων αυξήθηκε σε σχέση με τα τωρινά δεδομένα. Το δείγμα που μας δίνεται έχει μέγεθος 80 που σημαίνει πως αυτό του 2009 πιθανόν να ήταν μεγαλύτερο εφόσον οι πλειοψηφία των τιμών ήταν πιο μικρές. Αυτό προκύπτει καθώς γνωρίζουμε ότι ισχύει η Ιδιότητα Ασυμπτωτικής Κανονικότητας δειγματικού ποσοστού (το p^ προσεγγίζει την κανονική κατανομή καθώς το μέγεθος του δείγματος n αυξάνει).

c) Αρχικά πραγματοποιούμε διερευνητική ανάλυση των δεδομένων μας για να διαπιστώσουμε αν είναι κατάλληλα. Παρακάτω απεικονίζουμε το διάγραμμα barplot για τα δεδομένα μας, μέσω του οποίου βλέπουμε ότι τα δεδομένα μας είναι επιθυμητά, διότι το δείγμα είναι τυχαίο, δηλαδή πρόκειται για τυχαιοποιημένη και όχι αυτοματοποιημένη διαδικασία.



Όσον αφορά λοιπόν την αναλογία των χρωμάτων στα M&M's, παρατηρούμε ότι η αναλογία δεν είναι η ίδια, καθώς τα ποσοστά για κάθε χρώμα είναι

διαφορετικά σε σχέση με εκείνα του ερωτήματος a). Συγκεκριμένα τα ποσοστά των χρωμάτων στα M&M's είναι τα εξής:

```
> prop.table(t2)
data2
      BLUE      BROWN      GREEN        RED     YELLOW
0.16071429 0.17857143 0.08928571 0.21428571 0.35714286
> barplot(t2)
```

Ενώ, τα ποσοστά των χρωμάτων για τα smarties είναι τα παρακάτω:

```
> prop.table(t)
data
  BLUE  BROWN  GREEN    RED YELLOW
0.1875 0.2750 0.1000 0.2375 0.2000
> barplot(t)
```