Στατιστική στην Πληροφορική
2020-2021

Φέκα Αγγελική A.M:3140290
Σταυρουλάκη Μαρία A.M:3160168

2η σειρά ασκήσεων

Άσκηση 1η :

a) Αρχικά για να καταλάβουμε αν τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε, φτιάχνουμε το stemplot των δεδομένων που μας δίνονται.

Stemplot:
0 | 44444
0 | 55556688899
1 | 013
1 |
2 |
2 | 8

Επιπλέον θα πρέπει να αναφερθεί ότι τρόπος δειγματοληψίας που χρησιμοποιήθηκε είναι κατάλληλος για την εξαγωγή στατιστικών συμπερασμάτων, εφόσον πρόκειται για μια τυχαία επιλογή από τον πληθυσμό των αιτήσεων μιας ημέρας.

Όσον αφορά το stemplot, παρατηρούμε ότι υπάρχει η ατυπική τιμή 28, πράγμα το οποίο μπορεί να αποτελέσει ένδειξη για ένα μη κανονικά κατανεμημένου πληθυσμού. Από την θεωρία γνωρίζουμε ότι οι μέθοδοι συμπερασματολογίας που βασίζονται στην κατανομή t είναι ακριβείς σε κανονικούς πληθυσμούς. Όμως γνωρίζουμε ότι η ακρίβεια είναι αρκετά καλή και σε μη κανονικά κατανεμημένους πληθυσμούς, αν και μόνο αν το μέγεθος του δείγματος είναι αρκετά μεγάλο. Εδώ το μέγεθος του δείγματος είναι n = 20 όπου είναι μεγαλύτερο από το 15. Επομένως, μπορούμε να συμπεράνουμε ότι η ακρίβεια των μεθόδων κρίνεται καλή.

b) Γνωρίζουμε ότι ο τύπος για το διάστημα εμπιστοσύνης για τη μέση τιμή είναι:

$$\bar{x} \pm z_* \frac{\sigma}{\sqrt{n}}$$

Από τα δεδομένα μπορούμε να υπολογίσουμε τα ακόλουθα:
Δειγματικός μέσος όρος: $\bar{x}$ = 77.4 millisecond

Δειγματική τυπική απόκλιση: s = 55.52 millisecond
t_* = 2.093 για επίπεδο εμπιστοσύνης 95% και βαθμό ελευθερίας df = n -1 = 19

Επομένως με την χρήση του παραπάνω τύπου υπολογίζουμε το ζητούμενο διάστημα εμπιστοσύνης, το οποίο είναι [51.41365, 103.38635].

Άσκηση 2η :

a) Το λάθος βρίσκεται στην εφαρμογή του τύπου της τυπικής απόκλισης του δειγματικού μέσου, που είναι $\frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{20}}$.

b) Η μηδενική υπόθεση δεν μπορεί να χρησιμοποιηθεί σε μικρό δείγμα, άρα δε μπορούν να χρησιμοποιηθούν στη δειγματοληψία, παρά μόνο σε στατιστικά του γενικού πληθυσμού.

c) Για να απορριφθεί η μηδενική υπόθεση στην περίπτωση που αναφέρεται θα πρέπει το p-value να είναι μικρότερο από το βαθμό σημαντικότητας (p-value < α). Εδώ ο δειγματικός μέσος είναι μικρότερος άρα δεν θα πρέπει να απορριφθεί η H_0, εφόσον για να απορριφθεί η H_0 πρέπει ο δειγματικός μέσος να είναι αρκετά μεγαλύτερος του 54.

d) Θα πρέπει να επιλέξουμε μια οριακή τιμή, όπως το βαθμό σημαντικότητας, κάτω της οποίας θα απορρίπτουμε τη μηδενική τιμή και όχι να το κάνουμε αυθαίρετα.

Άσκηση 3η :

a) Με την βοήθεια του προγράμματος της R βρήκαμε ότι το p-value για την εναλλακτική υπόθεση $H_a : \mu > \mu_0$ είναι 1- F(z) = 0.09012267, όπου F(z) είναι η αθροιστική συνάρτηση κατανομής της z(n-1) και η τιμή του στατιστικού ελέγχου z = 1.34 σε έλεγχο σημαντικότητας με $H_0 : \mu = \mu_0$.

b) Με την βοήθεια του προγράμματος της R βρήκαμε ότι το p-value για την εναλλακτική υπόθεση $H_a : \mu < \mu_0$ είναι F(z) = 0.9098773, όπου F(z) είναι η αθροιστική συνάρτηση κατανομής της z(n-1) και η τιμή του στατιστικού ελέγχου z = 1.34 σε έλεγχο σημαντικότητας με $H_0 : \mu = \mu_0$.

c) Με την βοήθεια του προγράμματος της R βρήκαμε ότι το p-value για την εναλλακτική υπόθεση $H_a : \mu \neq \mu_0$ είναι 2F(-|z|) = 0.1802453, όπου F(z) είναι η αθροιστική συνάρτηση κατανομής της z(n-1) και η τιμή του στατιστικού ελέγχου z = 1.34 σε έλεγχο σημαντικότητας με $H_0 : \mu = \mu_0$.

Άσκηση 4η :

a) Ένας δίπλευρος έλεγχος υπόθεσης (δηλ. H_0 : μ = 30 ενάντια H_1 : μ != 30) με βαθμό σημαντικότητας α απορρίπτει την H_0 όταν το μ_0 βρίσκεται έξω από το διάστημα εμπιστοσύνης βαθμού C = 1 − α. Προκειμένου να έχουμε διάστημα εμπιστοσύνης C = 95% θα πρέπει 1 - α = 95%, άρα α = 5%. Η p-value είναι 0.04 = 4%. Έχουμε ότι p-value < α, επομένως απορρίπτεται και δεν περιέχεται στο διάστημα εμπιστοσύνης.

b) Στο διάστημα 90% θα έχουμε 1 - α = 90%, δηλαδή α = 10%, το οποίο πάλι απορρίπτεται αφού είναι ήδη μεγαλύτερο από την προηγούμενη τιμή απόρριψης.

Άσκηση 5η :

a) Το stemplot που προκύπτει από τα δεδομένα του *Πίνακα 1* είναι το εξής:

Stemplot:
0 | 6
2 |
4 | 459
6 | 578912233357
8 | 01233612

Από το Stemplot βλέπουμε πως υπάρχει ένα ατυπικό σημείο, το οποίο μάλλον οφείλεται σε λάθος εισαγωγή των στοιχείων, καθώς είναι αδύνατον ένα ενήλικο άτομο να ζυγίζει 6 κιλά. Για αυτό το λόγο θα αγνοήσουμε αυτό το δεδομένο. Το καινούργιο Stemplot που προκύπτει με τα νέα δεδομένα είναι το παρακάτω:

Stemplot:
5 | 459
6 | 5789
7 | 012233357
8 | 012336
9 | 12

Από το Stemplot που προκύπτει τώρα παρατηρούμε ότι η κατανομή του πληθυσμού είναι αρκετά συμμετρική. Επιπλέον, επειδή το μέγεθος των δεδομένων είναι n = 24 μπορούμε να εφαρμόσουμε την μέθοδο που βασίζεται στην κατανομή t με βαθμό ελευθερίας 23 (df = n -1 = 23).

Από τα δεδομένα μπορούμε να υπολογίσουμε τα ακόλουθα:
Δειγματικός μέσος όρος: $\bar{x}$ = 73.79

Δειγματική τυπική απόκλιση: s = 9.98
t_* = 2.069

Επομένως με την χρήση του τύπου για το διάστημα εμπιστοσύνης για τη μέση τιμή το ζητούμενο διάστημα εμπιστοσύνης είναι:
$\bar{} \pm t_* \frac{\sigma}{\sqrt{n}}$ = [69.57826, 78.00507]

b) Αρχικά φτιάχνουμε τα Stemplot για τα αγόρια και για τα κορίτσια αντίστοιχα:

Stemplot για τους άνδρες:
6 | 8
7 | 2233
7 | 57
8 | 013
8 | 6
9 | 12

Stemplot για τις γυναίκες:
5 | 459
6 | 579
7 | 013
8 | 23

Από τα Stemplot παρατηρούμε ότι κατανομές των δύο πληθυσμων είναι αρκετά συμμετρικές. Επιπλέον, επειδή το μέγεθος των συνολικών δεδομένων είναι n = 24, συμπεραίνουμε ότι η ακρίβεια είναι καλή και επομένως μπορούμε να προχωρήσουμε.

Από τα δεδομένα μπορούμε να υπολογίσουμε τα ακόλουθα:
Μέγεθος δεδομένων για τους άνδρες: n_m = 13
Δειγματικός μέσος όρος για άνδρες: $\bar{x}_m$ = 78.69 kg
Δειγματική τυπική απόκλιση για άνδρες: s_m = 7.6 kg

Μέγεθος δεδομένων για τις γυναίκες: n_f = 11
Δειγματικός μέσος όρος για γυναίκες: $\bar{x}_m$ = 68 kg
Δειγματική τυπική απόκλιση για γυναίκες: s_m = 9.57 kg

Επομένως με την χρήση της συνάρτησης t-test στην R προκύπτει ότι το 80% διάστημα εμπιστοσύνης είναι [5.948, 14.437] με df = 19.00547 και t_* = 1.328.

c) Προκειμένου να ελέγξουμε αν το κάπνισμα έχει σχέση με το βάρος, θα χρησιμοποιήσουμε τον δίπλευρο έλεγχο σημαντικότητας $H_0 : \mu_1 = \mu_2$, όπου μ_1 είναι το μέσο βάρος εκείνων που καπνίζουν και μ_2 το μέσο βάρος εκείνων που

δεν καπνίζουν.

Αρχικά παρατηρούμε ότι τα αποτελέσματα μας είναι κατάλληλα για την εξαγωγή συμπερασμάτων, καθώς είναι τυχαία δείγματα από τον πληθυσμό των καπνιστών και των μη καπνιστών, είναι αρκετά στο σύνολο (n=24) και οι κατανομές των πληθυσμών είναι αρκετά συμμετρικές, όπως βλέπουμε από τα παρακάτω Stemplots.

Stemplot βάρους καπνιστών:
5 | 9
6 | 5
7 | 137
8 | 0236
9 | 2

Stemplot βάρους μη καπνιστών:
5 | 45
6 | 789
7 | 022335
8 | 13
9 | 1

Επομένως μπορούμε να προχωρήσουμε στον έλεγχο. Από τα δεδομένα μπορούμε να υπολογίσουμε τα ακόλουθα:

Μέγεθος δεδομένων για τους καπνιστές: $n_1 = 10$
Δειγματικός μέσος όρος για τους καπνιστές: $\bar{x}_1 = 76.8$
Δειγματική τυπική απόκλιση για τους καπνιστές: $s_1 = 9.976$

Μέγεθος δεδομένων για τους μη καπνιστές: $n_2 = 14$
Δειγματικός μέσος όρος για τους μη καπνιστές: $\bar{x}_2 = 71.64$
Δειγματική τυπική απόκλιση για τους μη καπνιστές: $s_2 = 9.763$

Επομένως από τον γνωστό τύπο έχουμε:
$t = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{s_1^2}{n1} + \frac{s_2^2}{n2}}} = 1.2597$ με $df = \min\{n_1 - 1, n_2 - 1\} = 9$
Άρα p-value = 0.2395

Άσκηση 6η :

a) Έχουμε ένα τυχαίο δείγμα με συνολικά 20 δεδομένα τα οποία μας είναι αρκετά, έτσι ώστε να μπορέσουμε να πραγματοποιήσουμε τις μεθόδους συμπερασματολογίας. Επιπλέον παρατηρούμε από το stemplot ότι τα

δεδομένα μας δεν είναι σημαντικά μη συμμετρικά.

Το stemplot των τιμών:
4 | 6999
5 | 012334444
5 | 67
6 | 0334
6 | 9

Επομένως ως συμπέρασμα έχουμε ότι τα δεδομένα είναι κατάλληλα.

b) Μέση τιμή $\bar{x}$ = 5.5 και τυπική απόκλιση s = 0.6008766

c) 95% διάστημα εμπιστοσύνης = $\bar{} \pm t_* \frac{s}{\sqrt{n}}$ = [5.219, 5.781], όπου χρησιμοποιούμε t_* = 2.093, df = 19.

Άσκηση 7η :

Αρχικά θα πρέπει να υπολογιστούν οι διαφορές μεταξύ της εκτίμησης του συνεργείου και του εμπειρογνώμονα.

Αυτοκίνητο	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Διαφορά	100	50	-50	0	-50	200	250	200	150	300

Το stemplot των διαφορών είναι:
-0 | 55
 0 | 05
 1 | 05
 2 | 005
 3 | 0

Παρόλο που τα δεδομένα μας είναι μικρά σε αριθμό, φαίνονται αρκετά συμμετρικά και ακολουθούν την κανονική κατανομή. Επομένως μπορούμε να προχωρήσουμε.

Έστω μ είναι η μέση τιμή της διαφοράς, τότε ο έλεγχος σημαντικότητας που θα θεωρήσουμε είναι:

- Μηδενική υπόθεση: H_0: μ = 0
- Εναλλακτική υπόθεση: H_α = μ > 0

Θα χρησιμοποιήσουμε την εναλλακτική μ > 0, διότι μας ενδιαφέρει να διαπιστώσουμε εάν το συνεργείο υπερεκτιμά τις ζημίες, και όχι αν τις

υποεκτιμά.

Από τα δεδομένα μπορούμε να υπολογίσουμε τα ακόλουθα:

- Δειγματικός μέσος όρος $\bar{x}$ = 115
- Μέγεθος δεδομένων n = 10
- Δειγματική τυπική απόκλιση s = 124.8332220738267

Επομένως από τον γνωστό τύπο έχουμε: t = $\frac{\bar{x}}{\frac{s}{\sqrt{n}}}$ = 2.913

Άρα το p-value = 0.0086.
Το p-value είναι αρκετά μικρό και μας οδηγεί στο να απορρίψουμε τη μηδενική υπόθεση, δηλαδή το συνεργείο υπερεκτιμά τις ζημιές.

Άσκηση 8η :

a) Από τα Stemplot παρατηρούμε ότι κατανομές των δύο πληθυσμων είναι αρκετά συμμετρικές. Επιπλέον, επειδή το μέγεθος των συνολικών δεδομένων είναι n = 116, συμπεραίνουμε ότι η ακρίβεια είναι καλή και επομένως μπορούμε να προχωρήσουμε.

Stemplot για το ύψος των γυναικών:
15 | 4
15 | 688
16 | 00000122334
16 | 555556777889
17 | 0000
17 | 57788
18 |
18 | 55

Stemplot για το ύψος των ανδρών:
16 | 3579
17 | 000112333
17 | 44444455566
17 | 77888888999
18 | 00000000000
18 | 111222333333
18 | 44555555777
18 | 89
19 | 00134

Επομένως μπορούμε να προχωρήσουμε στον έλεγχο. Από τα δεδομένα μπορούμε να υπολογίσουμε τα ακόλουθα:
Μέγεθος δεδομένων για το ύψος των γυναικών : n_M = 38

Δειγματικός μέσος όρος για το ύψος των γυναικών: $\bar{x}_M$ = 1.666579
Δειγματική τυπική απόκλιση για το ύψος των γυναικών: s_M = 0.07419027sd

Μέγεθος δεδομένων για το ύψος των ανδρών : n_M = 78
Δειγματικός μέσος όρος για το ύψος των ανδρών: $\bar{x}_M$ = 4.085789
Δειγματική τυπική απόκλιση για το ύψος των ανδρών: s_M = 19.98295

Θα χρησιμοποιήσουμε βαθμό ελευθερίας t_* = 1.980808,μ με df = min{78, 38} = 38.

Επομένως με την χρήση του τύπου για το διάστημα εμπιστοσύνης για τη μέση τιμή το ζητούμενο διάστημα εμπιστοσύνης είναι:
95% Confidence Interval= $\overline{x_M} - \overline{x_F} \pm \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}}$ = [-2.165364, 7.003785].

b) Προκειμένου να απαντήσουμε στο ερώτημα θα θέσουμε κατάλληλη μηδενική συνθήκη.

Έστω μ_M ο μέσος όρος βαθμού στο μάθημα των Πιθανοτήτων των ανδρών και μ_F ο μέσος όρος βαθμού στο μάθημα των Πιθανοτήτων των γυναικών.

Έχουμε: $H_0 : \mu_M = \mu_F$ και $H_a : \mu_M > \mu_F$

Προκειμένου να δούμε αν οι άντρες φοιτητές που έχουν πάρει ή θα έπαιρναν το μάθημα «Στατιστική στην Πληροφορική» επιτυγχάνουν μεγαλύτερο μέσο βαθμό στο μάθημα των Πιθανοτήτων από τον αντίστοιχο πληθυσμό γυναικών, σε επίπεδο σημαντικότητας 5%. Θα πρέπει να μετρήσουμε την πιθανότητα p-value και να ισχύει το H_a , δεδομένο ότι χρησιμοποιούμε μηδενική υπόθεση. Στην περίπτωση που το p-value είναι μικρότερο του 5% η μηδενική υπόθεση θα απορριφθεί.

Με την βοήθεια του προγράμματος της R βρήκαμε ότι το p-value είναι:
p-value = 0.2403 > 5% .

c) Για να δούμε αν ο μέσος βαθμός στα Μαθηματικά 1 διαφέρει από το μέσο βαθμό στις Πιθανότητες, μεταξύ των φοιτητών που έχουν πάρει ή θα έπαιρναν το μάθημα «Στατιστική στην Πληροφορική» θα πρέπει να βρούμε το p-value για το εξής: $H_0 : \mu_1 = \mu_2$ και $H_a : \mu_1 \neq \mu_2$. Όπου μ_1 είναι ο μέσος όρος των φοιτητών στα Μαθηματικά 1 και μ_2 ο μέσος όρος των φοιτητών στις Πιθανότητες. Αν το p-value βγεί αρκετά χαμηλό, τότε θα πρέπει να απορρίψουμε την μηδενική υπόθεση.

Με την βοήθεια του προγράμματος της R βρήκαμε ότι το p-value είναι:
p-value = 0.5445