2η Σειρά ασκήσεων

Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής Μάθημα: Στατιστική στην Πληροφορική
Ακαδημαϊκό έτος: 2019–20

Κωνσταντινος Νικολουτσος → p3170122
Νικηφόρος Βλάχος → p3170018

Ασκηση 1)

a. Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε; Εξηγήστε.

Ο τρόπος δειγματοληψίας ειναι ιδανικός για τα συμπερασματα στατιστικών διότι ακολουθεί την ιδεα του SRS(Simple Random Samples).

Στην παρακάτω εικόνα βλέπουμε το stemplot μέσω της R για τα συγκεκριμενα δεδομενα.

```
> stem(sample, scale = 1, width = 80, atom = 1e-08)

  The decimal point is 2 digit(s) to the right of the |

  0 | 44444
  0 | 55556688899
  1 | 013
  1 |
  2 |
  2 | 8
```

Πιστεύουμε πως τα δεδομένα που αντλήσαμε θα μπορούσαν να είναι περισσότερα.

Παρόλα αυτα n>=15 επομένως δεν θα επηρεαστει τοσο πολύ το αποτελεσμα που θα βρουμε αν ο πλυθησμος δεν ειναι τοσο κανονικός.
Φυσικα αν ειχαμε λιγοτερα δεδομενα θα κρινόταν αναγκαίο να είχαν κανονική κατανομη!

b. Δώστε ένα 95% διάστημα εμπιστοσύνης για τη μέση τιμή του χρόνου διεκπεραίωσης.

Χρησιμοποιώντας την μεθοδολογία που βρίσκεται στις διαφάνειες προκυπτει ότι:

Αρχικά υπολογίζουμε τις εκτιμήτριες συναρτήσεις για τα ακόλουθα:

$\bar{X}$ = 77.4 millisecond → Δειγματικός μέσος όρος
s = 55.52 millisecond → Δειγματική τυπική απόκλιση(standard error)
$T*$ = 2.093 (Χρησιμοποιήσαμε βαθμο ελευθερίας 19 και α = 5%)

Οπότε τώρα θα χρησιμοποιήσουμε τον γνώστο τύπο που θα μας δώσει το ζητούμενο διαστημα εμπιστοσύνης (confidential interval)

- Για μια άγνωστη τυπική απόκλιση: $\left(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$

Τέλος προκύπτει ότι το ζητούμενο διάστημα είναι: **[51.41365, 103.38635]**

Ασκηση 2)

a. Λαμβάνεται ένα τυχαίο δείγμα μεγέθους 20 από πληθυσμό με τυπική απόκλιση 12. Η τυπική απόκλιση του δειγματικού μέσου είναι 12/20.

Απάντηση: Λάθος διότι γνωρίζουμε οτι η τυπική αποκλιση του δειγματικου μέσου(sampling mean) ειναι ίση με $\frac{\sigma}{\sqrt{n}}$ = $\frac{12}{\sqrt{20}}$

b. Ένας ερευνητής χρησιμοποιεί σε έναν έλεγχο σημαντικότητας τη μηδενική υπόθεση $H0$: $\overline{x}$= 10.

Απάντηση: Η μηδενική υπόθεση (null hypothesis) και γενικότερα οι υποθέσεις δεν μπορουν να βασίζονταιι πανω στην δείγματοληψία. Αντίθετα αφορά στατιστικά του γενικού πληθυσμού!

c. Σε μια στατιστική έρευνα με $\overline{x}$= 45 απορρίπτεται η μηδενική υπόθεση $H0$: μ = 54 όταν η εναλλακτική είναι Ha: μ > 54

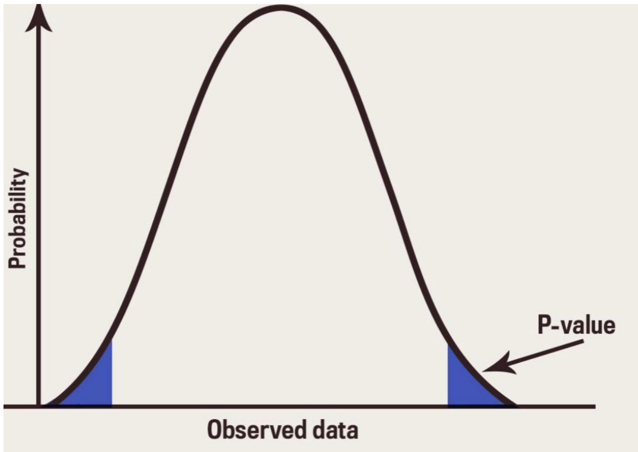
Απάντηση: Για να γίνει reject η μηδενική υπόθεση και έχουμε βλέψεις για την ενναλακτική θα πρέπει ο το p-value να είναι μικρότερο απο τον βαθμο σημαντικότητας (p-value < α).
Στην συγκεκριμένη περίπτωση ωστόσο ο δειγματικος μεσος ειναι μικρότερος αρα δεν θα πρεπει να απποριφθει η μηδενική υποθεση (απλη λογική)!

d. Σε μια στατιστική έρευνα όπου p value = 0.52 απορρίπτεται η μηδενική υπόθεση.

Απάντηση: Αν και 52% ειναι ενα μεγάλο ποσοστό, αυτο δεν ειναι σωστό καθως η απορριψη της μηδενικής υπόθεσης εξαρτάται απο το βαθμό σημαντικότητας. Αν λοιπόν α>0.52 τότε θα απορρίπταμε την αρχική υπόθεση και θα είχαμε βλέψεις για την ενναλακτική! Αξίζει να σημειώσουμε οτι τις περισσότερες φορές το α ειναι 0.1, 0.05, 0.01

Ασκηση 3)

Ο υπολογίσμός αυτός ουσιαστικά κρύβει την μέτρηση εμβαδού(βλεπε παρακάτω εικόνα)
Για να υπολογίσουμε τα παρακάτω βρήκαμε το z-table στην wikipedia!



a) Ποιο είναι το p value για την εναλλακτική υπόθεση Ha: μ > μ0;

p-value $\equiv$ **P(**$z \geq 1.34$**)** $\approx$**0.090122**

b) Ποιο είναι το p value για την εναλλακτική υπόθεση Ha: μ < $\mu 0$;

Με την ιδια λογική προκύπτει ότι **p-value = 0.909877**

c) Ποιο είναι το p value για την εναλλακτική υπόθεση Ha: $\mu \neq \mu 0$;

Εδώ έχουμε ενναλακτική υπόθεση διπλής κατεύθυνσης αλλα είναι το ίδιο διαδικάσία
p-value = 0.180245

Ασκηση 4)

Το p value για ένα δίπλευρο έλεγχο με μηδενική υπόθεση $H0$: μ = 30 είναι 0.04.

a. Η τιμή 30 περιέχεται στο 95% διάστημα εμπιστοσύνης για τη μέση τιμή μ; Γιατί;

Με μια γρήγορη ματιά παρατηρούμε οτι το επιπεδο σημαντικότητα ειναι α = 100% - 95% = 5%

Καθως επίσης pvalue = 4%. Αυτό σημαίνει ότι **pvalue < α** και αρα μπορούμε να κανουμε reject την μηδενική υπόθεση (null hypothesis). Επομένως λοιπον δεν ειμαστε σίγουροι αν ανήκει η τιμή 30 στο διαστημα εμπιστοσύνης(confidential interval)

b. Η τιμή 30 περιέχεται στο 90% διάστημα; Γιατί;

Φυσικα και όχι, το 30 δεν ανήκει στο confidential interval για τον ίδιο λόγο με το πανω ερώτημα.

Ασκηση 5)

Αρχικά θα πρέπει να αναφέρουμε οτι παρατηρήθηκε outlier(στην εισαγωγη δεδομενων μαλλον) διότι ειναι αδυνατο να υπαρχει άτομο που να ζυγίζει 6kg. Για αυτο τον λόγο θα το αγνοήσουμε απο τα δεδομένα μας για να μην αλλοιωθούν! Επίσης για n = 24 και για αγνωστο standard deviation μπορουμε να εφαρμόσουμε την μέθοδο που βασίζεται στην κατανομή t με βαθμο ελευθεριας 23.

a. Δώστε ένα 95% διάστημα εμπιστοσύνης για το μέσο βάρος των ενηλίκων κατοίκων Αθήνας.

$n = 24$

$\overline{x} = 73.79$

$s = 9.98 \rightarrow$ standard error

$df = 23$

$t_{*} = 2.069$

Απο τα παραπανω λοιπόν προκύπτει ότι: **CONFIDENCE_INTERVAL = [69.5782, 78.0050]**

b. Δώστε ένα 80% διάστημα εμπιστοσύνης για τη διαφορά του μέσου βάρους μεταξύ ανδρών και γυναικών (ενηλίκους κατοίκους Αθηνών).

Αρχικα ας ρίξουμε μια ματια στα δεδομένα δημιουργώνστας stemplot για τα αγορια και τα κοριτσια:

```
> stem(man, scale = 1, width = 80, atom = 1e-08)

  The decimal point is 1 digit(s) to the right
of the |

  6 | 8
  7 | 2233
  7 | 57
  8 | 013
  8 | 6
  9 | 12
```

```
> stem(woman, scale = 1, width = 80, atom =
1e-08)

  The decimal point is 1 digit(s) to the right
of the |

  5 | 459
  6 | 579
  7 | 013
  8 | 23
```

Παρατηρούμε οτι τα δεδομένα ειναι αρκετα συμμετρικά. Αρα προχωράμε κανονικά και έχουμε ότι:

m → συμβολίζει το male
f → συμβολίζει το female

$n_m = 13$

$n_f = 11$
$\overline{x_m} = 78.69kg$
$\overline{x_f} = 68kg$
$s_m = 7.6kg$
$s_f = 9.52kg$

Χρησιμοποιώντας την συνάρτηση t-test στην R προκύπτει ότι:
80%CONFIDENCE_LEVEL = [5.94, 14.43]

c. Το κάπνισμα έχει σχέση με το βάρος; Διατυπώστε έναν κατάλληλο έλεγχο σημαντικότητας και σχολιάστε τα ευρήματά σας.

Θα κάνουμε αυτο που διδαχθήκαμε. Θα δοκιμάσουμε τον δίπλευρο έλεγχο με:
$H_0 : \mu_{\nu\alpha\iota} = \mu_{o\chi\iota}$, οπου

Ναι → αυτη που καπνοιζουν	Οχι → αυτη που δεν καπνίζουν

(*Ισως να επρεπε να γίνει καποια έρευνα τωρα με τον αντι-καπνιστικό νομο 😅)

Πρώτα λοιπόν θα πρέπει να γίνει έλεγχος των δεδομενων για την καταλληλοτητα τους Παρατηρούμε λοιπον οτι τα αποτελέσματα ειναι ιδανική για εξαγωγή συμπερασμάτων αφου αποτελουν τυχαια δείγματα και ειναι και αρκετα για να να λειτουργήσουν καλα οι μεθοδοι συμπερασματολογιας που γνωρίζουμε.

```
> stem(smoke_yes, scale = 1, width = 80, atom =
1e-08)

  The decimal point is 1 digit(s) to the right
of the |

  5 | 9
  6 | 5
  7 | 137
  8 | 0236
  9 | 2
```

```
> stem(smoke_no, scale = 1, width = 80, atom =
1e-08)

  The decimal point is 1 digit(s) to the right
of the |

  5 | 45
  6 | 789
  7 | 022335
  8 | 13
  9 | 1
```

Εφαρμόζουμε τον γνωστό τύπο βρίσκουμε:

$$t = \frac{\overline{x_{yes}} - \overline{x_{no}}}{\sqrt{\frac{s_{yes}{}^2}{n_{yes}} + \frac{s_{no}{}^2}{n_{no}}}} = 1.2597 \quad \text{με df} = 9$$

P-value = 0.2395

Θα λέγαμε οτι δεν έχει τοσο σχέση το τσιγάρο με τα κιλά. Δεν μπορούμε να απορρίψουμε την μηδενική μας υπόθεση.

Ασκηση 6)

a. Τα δεδομένα είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε; Εξηγήστε.

```
> stem(sample, scale = 1, width = 80, atom =
1e-08)

  The decimal point is at the |

  4 | 6999
  5 | 012334444
  5 | 67
  6 | 0334
  6 | 9
```

Τα δεδομένα δεν φαίνονται να ειναι εξαιρετικά ασσύμετρα. Επίσης είναι επιλεγμένα στην τύχη σύμφωνα με την εκφώνηση. Τέλος έχουμε n=20 που ειναι αρκετά για να πραγματοποιήσουμε τις μεθόδους συμπερασματολογίας μας.

b. Βρείτε τη μέση τιμή και τυπική απόκλιση για τα δεδομένα αυτά.

```
> mean(sample)
[1] 5.5
> sd(sample)
[1] 0.6008766
```

c. Εκτιμήστε τη μέση τιμή μ της απόδοσης του αυτοκινήτου, με ένα 95% διάστημα εμπιστοσύνης χρησιμοποιώντας τα παραπάνω δεδομένα.

Θα χρησιμοποιήσουμε τον γνωστό τρόπο με την κατανομή t με βαθμο ελευθερίας 19 για να βρούμε το confidence interval.

95%CONFIDENCE_INTERVAL = $\overline{X} \pm \ t_{*} \frac{s}{\sqrt{n}} \ = \ = [5.219, \ 5.781]$

Ασκηση 7)

Σε αυτήν την περίπτωση δεν μπορούμε να εφαρμόσουμε τις μεθοδολογιες που γνωρίζουμε για ανεξαρτητα δείγματα διότι τα δειγματα μας εξαρτιουνται. Αυτο συμβαίνει διότι μια μεγάλη εκτίμηση ζημιας του εμπειρογνώμονα επηρεάζει το συνεργειο (εχει σχεση).

Θα φτίαξουμε εναν πίνακα διαφοράς συνεργείου με εμπειρογνώμονα:

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

Ας δούμε αν τα δεδομένα μας είναι κατάλληλα με την βοήθεια του stemplot.
Αν και έχουμε μικρό αριθμό δεδομένων φαίνονται αρκετα συμμετρικά και να ακολουθούν την κανονική κατανομή

```
> stem(difference, scale = 1, width = 80, atom
= 1e-08)

  The decimal point is 2 digit(s) to the right
of the |

  -0 | 55
   0 | 05
   1 | 05
   2 | 005
   3 | 0
```

Ας κάνουμε τους εξης ελεγχου σημαντικότητας:
Εστω μ να ειναι η μεση τιμη της διαφοράς

H_0: μ = 0 (null hypothesis)

H_a: μ > 0 (alternative hypothesis)

Βρίσκοντας το στατιστικό ελέγχου t και ολα τα απαραίτητα προκύπτει ότι :

P-value $\simeq$ 0.0086

Το P-value ειναι αρκετα μικρό για τα συνηθισμένα επιπεδα σημαντικότας. Οπότε μπορούμε να πούμε ότι απορρίπτουμε την αρχική υπόθεση. Αν και για να ειμασταν πιο σίγουροι θα επρεπε να πέρναμε περισσότερο sampling, αν και φαίνεται στο συγκεκριμένο παράδειγμα να προσεγγίζουμε την κανονική κατανομη (γνωστη και ως γκαουσιανή)

Ασκηση 8)

Επειδή τα δεδομένα ηταν αρκετα να τα γράψουμε με το χέρι, αρχικά κάναμε κάποιες μετατροπες σε java για να γίνουν στην μορφή που θέλουμε(data transform/extract)

a. Βρείτε ένα 95% διάστημα εμπιστοσύνης για τη διαφορά του μέσου ύψους μεταξύ ανδρών και γυναικών φοιτητών πληροφορικής του ΟΠΑ.

Αρχικά αν πρεπει να δούμε αν τα δεδομένα αυτά ειναι κατάλληλα για τις συμπερασματικές μεθοδολογίες μας. Τα δεδομένα δειχνουν συμμετρικά . Επίσης έχουμε αρκετά μεγάλο n.

```
> stem(man, scale = 1, width = 80, atom =
1e-08)

  The decimal point is 2 digit(s) to the left
of the |

  170 | 00000
  172 | 0000
  174 | 00000
  176 | 00000
  178 | 000000
  180 | 0000000
  182 | 000000
  184 | 00000
  186 | 00
  188 | 00
  190 | 00
```

```
> stem(woman, scale = 1, width = 80, atom =
1e-08)

  The decimal point is 1 digit(s) to the left
of the |

  15 | 3
  15 | 88
  16 | 0011344
  16 | 55577888
  17 | 000014
  17 | 667
  18 |
  18 |
  19 | 0
```

Θα χρησιμοποιήσουμε βαθμο ελευθερίας $t_* = 2.048$, df = min{49, 28} = 28

CONFIDENCE_INTERVAL = $\overline{x_1} - \overline{x_2} \pm \sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}$ = [0.1046134, 0.1365091]

b. Οι άνδρες φοιτητές πληροφορικής -που έχουν πάρει ή θα έπαιρναν το μάθημα «Στατιστική στην Πληροφορική»-, επιτυγχάνουν μεγαλύτερο μέσο βαθμό στο μάθημα των Πιθανοτήτων από τον αντίστοιχο πληθυσμό γυναικών; Απαντήστε σε επίπεδο σημαντικότητας 5%.

Θα απαντήσουμε στο ερώτημα αυτό θετοντας καταλληλη μηδενικη συνθήκη και κάνοντας τους γνωστούς υπολογισμούς.

Έστω μ_1 και μ_2 ο μεσος όρος βαθμου στις πιθανοτητες των αγορίων και κοριτσιων αντιστοιχα.

$$H_0 : \quad \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

Αρκει να μετρήσουμε την πιθανοτητα pvalue να ισχυει το H_a δεδομένο της μηδενικής υποθεσης.

Στην περίπτωση που το pvalue βγει μικρότερο απο 5% θα κανουμε reject την μηδενική υποθεση

Κάνοντας τους υπολογισμούς βγάλαμε ότι:

P-value = 0.024976 < 5% . Άρα κάνουμε έχουμε βλέψεις για το alternative hypothesis

(χρησιμοποιήσαμε βαθμό ελευθερίας ισο με το μικροτερο των δειγματων που ηταν των γυναικων.. Αυτο ήταν df=25)

c. 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»-;

Το συγκεκριμένο αφόρα διπλή κατεύθυνση και είναι παρόμοιο με το παραπάνω.
Πρέπει να βρούμε το pvalue για το εξης:

$$H_0 : \quad \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Οπου μ_1 , μ_2 είναι ο μεσος ορος εκεινων απο μαθηματικά1 και πιθανοτητες αντιστοιχα.
Αν το pvalue βγει αρκετα χαμηλό αυτό σημάινει οτι ισως θα πρεπει αν σκεφτουμε την απορριψη της null hypothesis. Πρακτικά αυτο θα σημαινει οτι αυτο που πηραμε απο τα δεδομενα ήταν πολυ σπανιο να παρθει δεδομενου οτι κανουμε SRS.