Ζέρβα Αθηνά-Μαρία : 3170207
Παπαγεωργίου Νικολέτα : 3170234

ΣΤΑΤΙΣΤΙΚΗ 2 ΣΕΙΡΑ ΑΚΣΗΣΕΩΝ

1.Μας ενδιαφέρει ο μέσος χρόνος διεκπεραίωσης μιας αίτησης (query) σε μια βάση δεδομένων. Από τις αιτήσεις μιας ημέρας, επιλέχθηκαν 20 τυχαίες και μετρήθηκαν οι ακόλουθοι χρόνοι διεκπεραίωσης σε milliseconds:

82	55	58	94	86	45	42	36	41	130
284	96	39	107	52	54	45	81	83	38

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

Το stemplot των τιμών είναι :

0|44444

0|55556688899

1|013

1|

2|

2|8

Εξαιτίας της ατυπικής τιμής,28, θα μπορούσε να θεωρηθεί ο πληθυσμός ως μη κανονικά καταναμεμήνος. Παρόλο που η συμπερασματολογία είναι ακριβείς σε κανονικά καταναμημένους πληθυσμούς ,το μεγάλο μέγεθος του δείγματος ευνοεί κι έτσι τα δεδομένα κρίνονται κατάλληλα.

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

Το διάστημα είναι : [51.42,103.38]

Που προκύπτει από τον τύπο $\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$

Όπου $\bar{x}$= 77.4, σ = 55.52 και $z*$ = 2.093

2. Παρακάτω δίνονται διάφορες περιπτώσεις λάθος εφαρμογής ελέγχων σημαντικότητας. Εξηγήστε για κάθε περίπτωση ποιο είναι το λάθος και γιατί είναι λάθος.

a. Η δειγματική τυπική απόκλιση s δεν είναι αμερόληπτος εκτιμητής. Άρα η s είναι ίση με $\sigma/\sqrt{n}$ → $12/\sqrt{20}$,όπου σ τυπική απόκλιση πληθυσμού και n μέγεθος δείγματος.

b. Ο έλεγχος σημαντικότητας με μηδενική υπόθεση έχει ως παράμετρο την μέση τιμή πληθυσμού μ και όχι τον μέσο όρο δείγματος $\bar{x}$.

c. Δεν θα έπρεπε να απορριφθεί η μηδενική υπόθεση , αφού ο μέσος όρος επιβεβαιώνει την μέση τιμή πληθυσμού.

d. Δεν αρκούν τα στοιχεία για να πούμε αν απορρίπτεται η υπόθεση ή όχι. Αν είχαμε το α (το οποίο είναι ο βαθμός σημαντικότητας) θα κάναμε α>0.52 και τότε μάλλον όντως θα απορρίπταμε την υπόθεση.

3. Σε έλεγχο σημαντικότητας με $H0$: $\mu = \mu 0$ η τιμή του στατιστικού ελέγχου z είναι 1.34.

a. Ποιο είναι το p value για την εναλλακτική υπόθεση Ha: $\mu > \mu 0$;

p value = 0.09012267

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

p value = 0.9098773

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

p value = 0.1802453

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;

Επίπεδο σημαντικότητας α = 100% - 95% = 5% και p value = 4%. Άρα p value < α , οπότε μπορούμε να πούμε ότι η μηδενική υπόθεση απορρίπτεται. Άρα δεν μπορούμε να πούμε με σιγουριά αν ανήκει η **: μ = 30** στο διάστημα εμπιστοσύνης.

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

Ισχύει ό,τι και στο παραπάνω ερώτημα. Δεν ανήκει στο διάστημα.

5. Θεωρήστε τα δεδομένα που δίδονται στον Πίνακα 1 παρακάτω, τα οποία προήλθαν από απλή τυχαία δειγματοληψία (simple random sampling – SRS) 25 ενηλίκων κατοίκων Αθήνας.

Πίνακας 1

Α/Α	ΦΥΛΟ	ΚΑΠΝΙΣΤΗΣ	ΒΑΡΟΣ
1	Α	ΝΑΙ	80
2	Α	ΟΧΙ	81
3	Α	ΟΧΙ	75
4	Α	ΝΑΙ	83
5	Γ	ΝΑΙ	71
6	Α	ΝΑΙ	73
7	Γ	ΝΑΙ	65
8	Γ	ΟΧΙ	67
9	Γ	ΟΧΙ	54
10	Α	ΝΑΙ	77
11	Γ	ΟΧΙ	55
12	Γ	ΟΧΙ	83
13	Α	ΟΧΙ	91
14	Γ	ΟΧΙ	6
15	Α	ΝΑΙ	92
16	Α	ΝΑΙ	86
17	Γ	ΟΧΙ	73
18	Γ	ΝΑΙ	82
19	Γ	ΟΧΙ	69
20	Α	ΟΧΙ	73
21	Γ	ΟΧΙ	70
22	Γ	ΝΑΙ	59
23	Α	ΟΧΙ	68
24	Α	ΟΧΙ	72
25	Α	ΟΧΙ	72

Στις απαντήσεις των παρακάτω ερωτημάτων, εξετάστε επίσης την καταλληλότητα των εκάστοτε δεδομένων.

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

[69.5281,78.0539]

Με $\bar{x} = 73.79167$, σ=9.978146 , z * = 2.093024

Αφαιρέσαμε την τιμή 14 καθώς δεν είναι δυνατό ένας άντρας ή μια γυναίκα να έχει βάρος 6 κιλά.

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

Εφόσον τα δεδομένα προέρχονται από σημαντικά μη συμμετρικές κατανομές ,έχουμε :

Από τον τύπο : $\overline{x_1} - \overline{x_2} \pm Z * \sqrt{\frac{\sigma_{1\hat{}2}}{n_1} + \frac{\sigma_{2\hat{}2}}{n_2}}$

[5.789,15.595]

Με n(m)= 13 , $\bar{x}(m) = 78.69231$, σ(m) = 7.598077

n(w) =11 , $\bar{x}(w) = 68$,σ(w) = 9.570789

Z* = 1.372

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

Πραγματοποιώντας τον δίπλευρο έλεγχο σημαντικότητας ,Ho : μ1 = μ2,ελέγχουμε αν το μέσο βάρος ,μ1 , των καπνιστών διαφέρει από το μ2 των μη καπνιστών .

Εφόσον και εδώ τα δεδομένα προέρχονται από σημαντικά μη συμμετρικές κατανομές ,έχουμε :

n1=10, $\overline{x_1} = 76.8$, σ1 = 9.975526

n2 = 14, $\overline{x_2} =$71.64286 , σ2= 9.76341

Τιμή στατικού ελέγχου : 1.2567

Pvalue = 0.2395

6. (Άσκηση 7.24 από "Introduction to the Practice of Statistics", Moore, McCabe, Craig, WH Freeman, 6th edition (IPS))

a.Τα δεδομένα έχουν επιλεγεί από τυχαία επιλογή και το δείγμα είναι n=20, άρα αρκετό για τις μεθόδους συμπερασματολογίας που γνωρίζουμε. Συμμετρικά →

4|6 9 9 9

5|7 2 4 6 4 3 1 0 4 3 4

6|4 3 9 0 3

b.
```
> mean(data)
[1] 5.5
 > sd(data)
[1] 0.6008766
```

c. Χρησιμοποιούμε την κατανομή t για να διάστημα εμπιστοσύνης C.

C= $\bar{x} \pm$ t * $\frac{s}{\sqrt{n}}$, και βαθμοί ελευθερίας n-1=19.

```
> t<- -qt(0.025,df=19)
> t
[1] 2.093024
> mean(data)+ c (-1,1) * t * sd(data)/sqrt(length(data))
[1] 5.218781 5.781219
```

7. (Άσκηση 7.34 από IPS) Μια ασφαλιστική εταιρία αυτοκινήτων συνεργάζεται με ένα συνεργείο αυτοκινήτων για εκτίμηση ζημιών όπου όμως υπάρχει η ανησυχία ότι το συνεργείο υπερεκτιμά τις ζημιές. Για να δει η εταιρία αν όντως συμβαίνει αυτό, έλαβε τις εκτιμήσεις ζημιών για 10 αυτοκίνητα από το εν λόγω συνεργείο και τις συνέκρινε με τις εκτιμήσεις ενός άλλου εμπειρογνώμονα. Ακολουθούν τα αποτελέσματα:

Αυτοκίνητο	1	2	3	4	5	6	7	8	9	10
Συνεργείο	500	1550	1250	1300	750	1000	1250	1300	800	2500
Εμπειρογνώμονας	400	1500	1300	1300	800	800	1000	1100	650	2200

Ελέγξτε τη μηδενική υπόθεση ότι δεν υπάρχει υπερεκτίμηση ζημιών από το συνεργείο. Τι συμπέρασμα βγάζετε;

Θα ελέγξουμε τη μέση τιμή του συνεργείου με τη μέση τιμή του εμπειρογνώμονα .

Θα ασχοληθούμε με τις διαφορές των δύο εκτιμήσεων . Οπότε το stemplot των διαφορών είναι το εξής :

-0|5 5

0|0 5

1|0 5

2|0 5 0

3|0

Παρόλο που το δείγμα είναι μικρό δεν αποκλείεται να είναι κοντά στη κανονική .

Με Ha : μ > 0 για να βρούμε αν υπάρχει υπερεκτίμηση ζημιών:

Pvalue= 0.008611

Με n =10 , $\bar{x}$=115 , σ = 124.8332 , z = 2.913

Τελικά συμπεραίνουμε από το μικρό pvalue πως υπάρχει υπερεκτίμηση από το συνεργείο κι έτσι απορρίπτουμε τη μηδενική υπόθεση.

8. Χρησιμοποιήστε τα δεδομένα με τις απαντήσεις του ερωτηματολογίου 2021 για να διατυπώσετε ελέγχους σημαντικότητας για τα ακόλουθα ερωτήματα:

a. Το δείγμα των δεδομένων είναι αρκετά μεγάλο και συμμετρικό (το έχουμε μελετήσει από την προηγούμενη άσκηση).

```
> t<- table(sex)
> t
sex
 F  M  O
34 86  3
> t<- - qt(0.025, df=33)
> t
[1] 2.034515
> t2<- -qt(0.025, df=85)
> t2
[1] 1.988268
```

Έχουμε df= min{34-1,86-1}=33, άρα t=2.034515.

C=$\overline{(x_1} - \bar{x}_2) * t * \pm\sqrt{\frac{s_1^2}{n_1}\frac{+s_2^2}{n_2}}$

```
>hm <- height[sex=="M"]
> mean(hm)
[1] 1.804265
> hf<- height[sex=="F"]
> mean(hf)
[1] 1.645333
>sd(hm)
[1] 0.07320623
> sd(hf)
```

[1] 0.08609271

> mean(hm)-mean(hf) + c (-1,1) *t*sqrt((sd(hf)^2 /(length(hf)))+(sd(hm)^2 /(length(hm))))

[1] 0.1222042 0.1956585

C=[0.1222042 , 0.1956585]

b. Έστω μ1 και μ2 ο μέσος όρος βαθμού στις πιθανότητες των αγοριών και κοριτσιών αντιστοιχία.

H0 : μ1=μ2

Hα: μ1>μ2

Αρκεί να μετρήσουμε την πιθανότητα p value , για να την συγκρίνουμε με το επίπεδο σημαντικότητας α. Εάν το p value βγει μικρότερο από 5% απορρίπτουμε την μηδενική υπόθεση.

Θα χρησιμοποιήσουμε τον βαθμό ελευθέριας των κοριτσιών, διότι όπως υπολογίσαμε στο προηγούμενο υποερώτημα, επιλέγουμε το ελάχιστο.

> pm<-prob[sex=="M"]

> pf<-prob[sex=="F"]

> t.test(pm,pf)

Welch Two Sample t-test

data: pm and pf

t = 1.0147, df = 57.807, **p-value = 0.3145**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.4406833 1.3465656

sample estimates:

mean of x mean of y

6.102941 5.650000

→ p-value=0.3145 <5% , οπότε μπορούμε να πούμε ότι η μηδενική υπόθεση απορρίπτεται.

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 αφορά διπλή κατεύθυνση και είναι παρόμοιο με το παραπάνω. Πρέπει να βρούμε το pvalue για το εξής:

H0 : μ1=μ2

Hα: μ1!=μ2

,όπου μ1 και μ2 είναι ο μέσος όρος εκείνων από μαθηματικά1 και πιθανότητες αντίστοιχα. Αν το p value βγει αρκετά χαμηλό αυτό σημαίνει ότι ίσως θα πρέπει να σκεφτούμε την απόρριψη της μηδενικής υπόθεσης. Πρακτικά αυτό θα σημαίνει ότι αυτό που πήραμε από τα δεδομένα ήταν πολύ σπάνιο να παρθεί δεδομένου ότι κάνουμε SRS.

> t.test(math&prob)

One Sample t-test

data: math & prob

t = 43.818, df = 100, **p-value < 2.2e-16**

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

0.9074588 0.9935313