Име и презиме

Број индекса

БИОСТАТИСТИКА И АНАЛИЗА ПОДАТАКА - ПРИМЕР ПИСМЕНОГ ИСПИТА

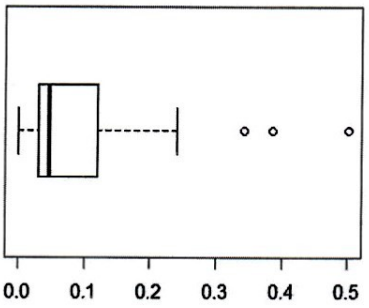
Питања

Напомена: Потребно је само уписати или заокружити тачан одговор.

1. Узорачка средња вредност $\bar{x}$ је
   (А) параметар популације (Б) параметар узорка (В) статистика

2. На слици је приказан боксплот одређених података. Расподела тих података је?
   (А) симетрична (Б) померена улево (В) померена удесно

Слика 1:



3. У кутији се налази 100 куглица, од којих је 25 плавих. Закључак да је вероватноћа да случајно изабрана куглица буде плава једнака $1/4$ добијена је коришћењем класичне дефиниције вероватноће.

4. Ако је 90% интервал поверења за непознати параметар μ једнак (1.2,3.9), то значи имамо поверење од 90% да је $1.2 \leq \mu \leq 3.9$, јер верујемо да је наш узорак један од оних 90% узорака на основу којих добијени интервал садржи μ.
   (А) да (Б) не

5. Уколико одбацимо исправну нулту хипотезу, правимо грешку I врсте.

6. Ако је коефицијент корелације између случајних величина X и Y једнак -0.9, онда су X и Y
   (А) некорелисане (Б) слабо линеарно корелисане (В) јако линеарно корелисане

7. Приликом тестирања хипотезе $H_0 : \mu = 0$ против алтернативе $H_1 : \mu < 0$, када се σ^2 оцењује, добијена је вредност тест статистике t_0. p-вредност овог теста рачуна се као
   (А) површина лево од t_0 (Б) површина десно од t_0 (В) двострука површина лево од t_0

8. Приликом испитивања повезаности променљивих X и Y формирамо табеле контигенције тако што фиксирамо број вредности за X у свакој од њених категорија. Хи-квадрат тест којим вршимо тестирање у том случају је
   (А) тест независности (Б) тест хомогености

9. Нека су A и B догађаји такви да је $P(A) = 0.6$ и $P(B|A) = 0.3$. Колика је вероватноћа $P(A \cap B)$?
   (А) 0.5 (Б) 0.9 (В) 0.6 (Г) 0.18

10. Који ће од следећих интервала поверења за непознати проценат p имати најмању дужину?
    (А) 90% интервал за $n = 100$ (Б) 95% интервал за $n = 100$
    (В) 90% интервал за $n = 144$ (Г) 95% интервал за $n = 81$

Задаци

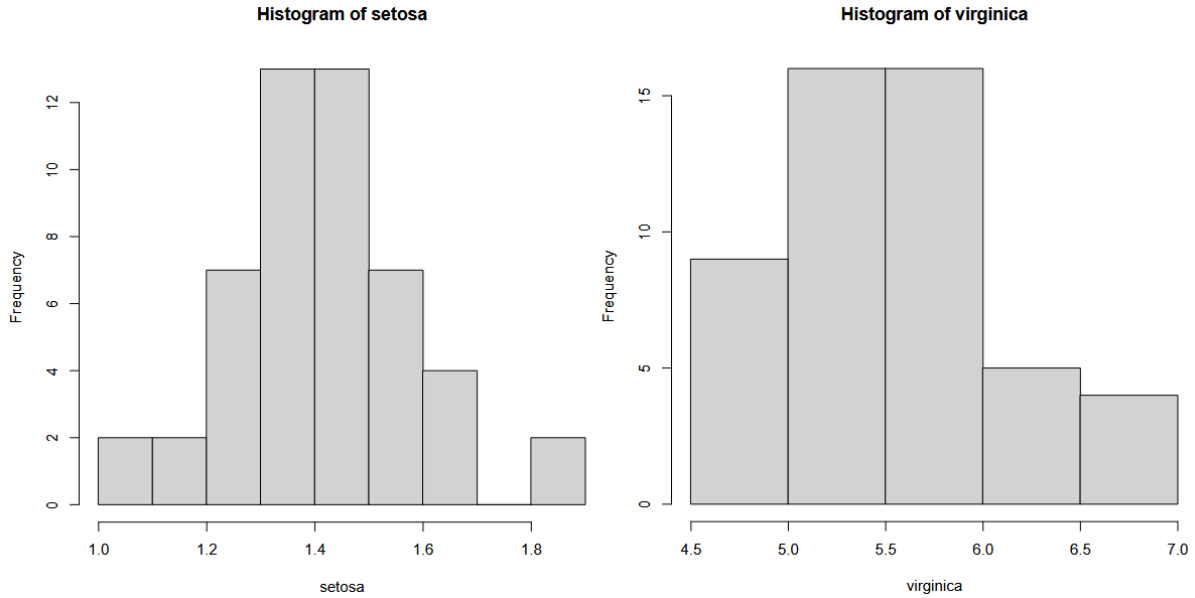
1. У датотекама *virginica.txt* и *setosa.txt* се налазе узорци о дужинама латица две различите врсте перуника. Напомена: Учитавање спољних датотека није рађено на часовима, већ је објашњен процес на самом тесту.

```
virginica = scan("virginica.txt")
setosa = scan("setosa.txt")
```

а) Нацртати хистограме за обе врсте цвета појединачно.

```
hist(virginica)
hist(setosa)
```

Слика 2:



б) Да ли подаци прате расподелу у облику звона? Да ли је расподела података асиметрична за било коју од две врсте? Ако јесте, у ком смеру је асиметрична?

Расподела дужина латица за врсту virginica је асиметрична удесно, док је расподела за врсту setosa симетрична и у облику звона.

в) Која од две врсте има дуже латице у просеку?

На основу хистограма се јасно види да virginica има дуже латице у просеку.

г) Израчунати дисперзије узорака. За коју врсту су распршеније вредности дужина латица?

```
var(virginica)

## [1] 0.3045878

var(setosa)

## [1] 0.03015918
```

Узорачка дисперзија за врсту virginica је око 10 пута већа, па су распршеније вредности дужине латице за њу.

2. Врши се истраживање о ефективности две вакцине, вакцине А и вакцине Б. У истраживању учествује 30 испитаника. Испитаници су подељени у 3 групе од по 10 испитаника - контролну групу, групу која прима вакцину А и групу која прима вакцину Б.

а) На колико начина је могуће распоредити испитанике у ове 3 групе?

$$\binom{30}{10}\binom{20}{10}$$

б) Истраживање траје 20 дана. Сматра се да испитаник контролне групе сваког дана има вероватноћу 0.01 да се инфицира. Колика је вероватноћа да на крају истраживања није још увек био инфициран? *Вероватноћа да се не инфицира у једном дану је* $q = 1 - 0.01 = 0.99$*, па је вероватноћа да се не инфицира 20 дана за редом једнака* $p = q^{20} = (0.99)^{20} = 0.818$.

```
(0.99)^20

## [1] 0.8179069
```

в) Колико очекујемо да је пацијената из контролне групе било заражено током истраживања? У контролној групи је 10 пацијената, и сваки има вероватноћу $1 - 0.818 = 0.182$ *да се зарази. Број заражених* X *тада прати биномну расподелу* $X \sim B(10, 0.182)$*. Очекивани број заражених у контролној групи је* $EX = 10 \cdot 0.182 = 1.82$.

3. Студенти Биолошког факултета су до сад имали два теста из Биостатистике. Дат је узорак резултата 10 студената који су изашли на оба теста. Постоји мишљење да је други тест био тежи од првог.

Тест 1	19	15	11	17	19	16	18	14	17	15
Тест 2	18	17	12	14	16	10	18	17	10	13

```
test1 = c(19, 15, 11, 17, 19, 16, 18, 14, 17, 15)
test2 = c(18, 17, 12, 14, 16, 10, 18, 17, 10, 13)
```

а) Одредити тачкасту оцену средње вредности поена за први и за други тест.

```
mean(test1)

## [1] 16.1

mean(test2)

## [1] 14.5
```

б) Формулисати нулту и алтернативну хипотезу којим би се испитала валидност наведеног мишљења.

Алтернативна хипотеза је да је други тест тежи од првог, што значи да је просечан број поена на њему мањи.

$$H_0 : m_1 \leq m_2$$
$$H_1 : m_1 > m_2$$

в) Тестирати хипотезу наведену под б) при нивоу значајности $\alpha = 5\%$. Можемо ли се сложити с наведеним мишљењем? Претпоставити да број поена на тесту прати нормалну расподелу.

Користимо Студентов тест за спарене узорке. Подаци су спарени зато што су за сваког од 10 студената дати поени на оба теста, а Студентов тест можемо да користимо јер је претпостављена нормалност података.

```r
t.test(test1, test2, alternative = 'greater', paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  test1 and test2
## t = 1.546, df = 9, p-value = 0.07826
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.297171       Inf
## sample estimates:
## mean of the differences
##                     1.6
```

p-*вредност је 0.07826, што је више од нивоа значајности, па немамо довољно доказа да прихватимо алтернативну хипотезу, односно не можемо да се сложимо с мишљењем да је други тест тежи од првог.*

4. Сматра се да сорта трешања *GraceStar* има већи плод у просеку од сорте *Burlat*. Да би се ово показало, посађено је по једно стабло обе сорте на 8 различитих плацева и следеће сезоне су измерени просечни дијаметри плодова (у мм).

Grace Star	22	13	15	20	19	22	19	23
Burlat	17	15	14	18	14	18	16	15

```r
grace_star = c(22, 13, 15, 20, 19, 22, 19, 23)
burlat = c(17, 15, 14, 18, 14, 18, 16, 15)
```

а) Да ли су ови узорци спарени? *Узорци су спарени, зато што су праћени резултати за обе сорте на сваком од 8 плацева.*

б) Поставити нулту и алтернативну хипотезу.

Нека је m_G *просечан дијаметар плода сорте* *GraceStar*, а m_B сорте *Burlat*.

$$H_0 : m_G \leq m_B$$

$$H_1 : m_G > m_B$$

в) Тестирати хипотезе тестом знакова са нивоом значајности $\alpha = 0.05$. Чему је једнака и шта представља вредност тест статистике?

```r
test_statistika = sum(grace_star > burlat)
test_statistika
```

```
## [1] 7
```

```r
p_vrednost = 1 - pbinom(test_statistika - 1, 8, 0.5)
p_vrednost
```

```
## [1] 0.03515625
```

Вредност тест статистике је 7, што представља број плацева на којим је плод сорте *GraceStar* *у просеку био већи.*

г) Извести закључке на основу резултата теста. p-*вредност је испод нивоа значајности, па можемо да закључимо да је плод сорте* $GraceStar$ *заиста већи.*

Напомена: p-*вредност је израчуната као вероватноћа, под нултом хипотезом, да се добије вредност тест статистике која је већа или једнака 7, што је израчунато као вероватноћа да се добије вредност већа од 6. Тест је било могуће и аутоматски извршити функцијом* $SIGN.test()$ *из пакета* $BSDA$.