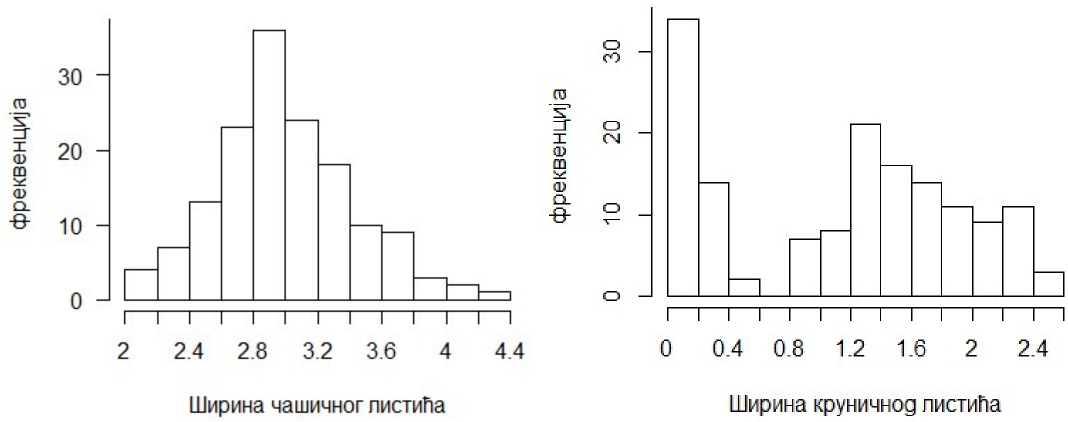
БИОСТАТИСТИКА - Писмени испит **1. јул 2018.**

1. [**12 поена**] Приказани су хистограми ширине (у цм) чашичних и круничних листића 150 цветова ириса. На основу датих хистограма проценити:

Слика 1:



а) [3 поена] Да ли су у просеку шири чашични или крунични листићи?

б) [3 поена] Да ли су у питању хистограми фреквенција или релативних фреквенција? Образложити одговор.

в) [2 поена] Ако бисте желели да покажете да је ширина круничног листића мања од ширине чашичног, који тест бисте користили ако су вредности узете на различитим цветовима?

г) [4 поена] Колики проценат чашичних листића има ширину између 2.4 и 4.4 центиметра? А колики проценат круничних листића има ширину између 2.4 и 4.4 центиметра?

Решење.

а) С хистограма видимо да је просечна ширина чашичних листића близу 3.0, а круничних близу 1, па закључујемо да су чашични листићи у просеку шири.

б) С обзиром да се на y-оси налази број листића који имају одговарајуће ширине, у питању је хистограм фреквенција. Хистограм релативних фреквенција на y-оси има удео (проценат).

в) Како немамо индиција да листићи немају нормалну расподелу, а знамо да су листићи узети с различитих цветова, користили бисмо T-тест на независним узорцима.

г) С хистограма видимо да ширину између 2.4 и 4.4 имају сви чашични листићи из узорка изузев прва два стуба у којима има $4 + 7 = 11$ елемената, па је тражени проценат $1 - \frac{11}{150} = 0.93 = 93\%$. Што се тиче круничних, у овом распону налазе се само они из последњег стуба, у којој их има 3, па је тражени проценат $\frac{3}{150} = 0.02 = 2\%$.

2. [**12 поена**] У популацији 1% жена има рак дојке. Уколико рак постоји, мамограм ће бити позитиван с вероватноћом 0.8. Уколико рак не постоји, мамограм ће бити негативан с вероватноћом 0.904.

а) [*6 поена*] Колика је вероватноћа да мамограм случајно изабране жене те популације буде позитиван?

б) [*6 поена*] Уколико је мамограм позитиван, колика је вероватноћа да жена стварно има рак дојке?

Решење.

Нека је R догађај да жена има рак, а M догађај да је мамограм позитиван. Знамо да је $P(R) = 0.01$, $P(M|R) = 0.8, P(\bar{M}|\bar{R}) = 0.904$. Лако рачунамо да је $P(\bar{M}|\bar{R}) = 1 - P(\bar{M}|\bar{R}) = 1 - 0.904 = 0.096$.

а) Користећи формулу потпуне вероватноће добијамо

$$P(M) = P(R)P(M|R) + P(\bar{R})P(M|\bar{R}) = 0.01 \cdot 0.8 + 0.99 \cdot 0.096 = 0.10304.$$

б) Користећи резултат под а) и формулу условне вероватноће, добијамо

$$P(R|M) = \frac{P(R)P(M|R)}{P(M)} = \frac{0.01 \cdot 0.8}{0.10304} = 0.078.$$

3. [**12 поена**] У циљу упоређивања два режима исхране за товљење јунади, девет парова животиња изабрано је из крда, тако да животиње у оквиру једног пара буду што сличнијих генетских карактеристика. Затим је један члан сваког пара подвргнут једном режиму, а други другом режиму исхране. Резултати представљају пораст тежине (у кг) након четири месеца.

режим I	298	211	262	227	269	276	239	282	278
режим II	249	230	234	229	265	241	264	299	228

а) [*3 поена*] Шта треба узети за нулту, а шта за алтернативну хипотезу?

б) [*6 поена*] С прагом значајности 0.05 извршити одговарајуће тестирање.

в) [*3 поена*] Који закључак се може извести на основу резултата теста?

Решење.

а) Нека је X пораст тежине приликом режима I, а Y приликом режима II и нека су им средње вредности редом μ_X и μ_Y. С обзиром да нам је циљ истраживања да упоредимо два режима исхране, нулта хипотеза је да нема разлике у средњој вредности пораста тежине ($H_0 : \mu_X = \mu_Y$), а алтернативна је да та разлика постоји ($H_1 : \mu_X \neq \mu_Y$).

б) Пошто су нам подаци дати у паровима, а немамо назнака да немају нормалну расподелу, користимо спарени T-тест.

Из узорка добијамо

$\bar{d} = \frac{1}{n}\sum(x - y) = \frac{1}{9}((298 - 249) + (211 - 230) + \cdots + (278 - 228)) = 11.44,$

и $s_D^2 = \frac{1}{n-1}\sum d^2 - \frac{n}{n-1}\bar{d}^2 = 878.5296.$

Вредност тест статистике је

$t_0 = \frac{\bar{d}}{s_D}\sqrt{n} = \frac{11.44}{29.64}\sqrt{9} = 1.16.$

Како је тест двострани, то p-вредност рачунамо као двоструку површину десно од вредности t_0. Из таблице Студентове расподеле за $\nu = 8$ видимо да се 1.16 налази између 1.10815 и 1.39682, па је површина десно од 1.16 између 0.10 и 0.15. Стога добијамо да је p-вредност теста између 0.2 и 0.3.

в) Како је p-вредност теста већа од 0.05, немамо довољно доказа да закључимо да постоји разлика у средњим вредностима пораста тежине код ова два режима исхране.

4. [**12 поена**] Мерена је дужина крила Y (у цм) код врабаца различите старости x (у данима). Резултати су приказани у следећој табели.

x	3.0	4.0	5.0	6.0	8.0	9.0	10.0	11.0	12.0	14.0	15.0	16.0	17.0
Y	1.4	1.5	2.2	2.4	3.1	3.2	3.2	3.9	4.1	4.7	4.5	5.2	5.0

а) [*1 поен*] Скицирати тачкасти график података. Може ли се наслутити линеарна зависност?

б) [*2 поена*] Одредити оцену праве линеарне регресије $m_{Y|x}$ методом најмањих квадрата и доцртати је на график. Слаже ли се с подацима?

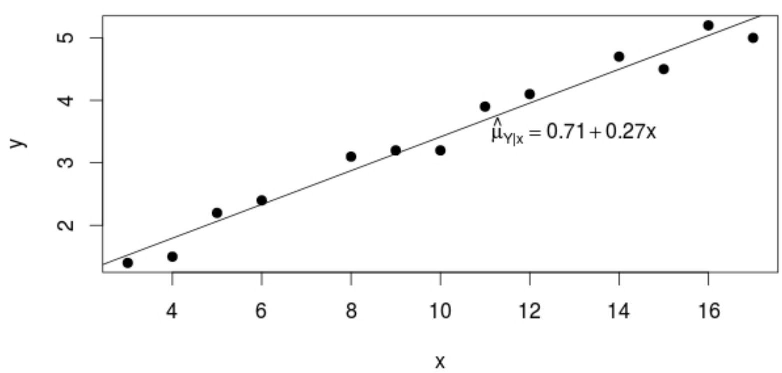
в) [*6 поена*] Одредити 90% интервал поверења за средњу дужину крила врапца старости 13 дана. Можемо ли на основу њега закључити да је средња дужина крила врапца те старости једнака 4.5?

г) [*3 поена*] Оценити средњу дужину крила врапца старог 60 дана.

[*бонус*] Ако бисте желели да тестирате да ли линеарни модел добијен под б) има смисла, шта бисте узели за нулту, а шта за алтернативну хипотезу? Одговор образложити.

Решење 1. а) Тачкасти график података приказан је на слици 2. Може се уочити линеарна зависност.

Слика 2:



б) На основу узорка моземо израчунати

$S_{yy} = \sum y^2 - n\bar{y}^2 = 19.66$, $S_{xx} = \sum x^2 - n\bar{x}^2 = 262$ и $S_{xy} = \sum xy - n\bar{x}\bar{y} = 70.8$.

Добијамо да су оцене параметара регресије $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 0.27$ и $\hat{\alpha} = \bar{Y} - \hat{\beta}x = 3.41 - 0.27 \cdot 10 = 0.71$. Одатле добијамо $\hat{\mu}_{Y|x} = 0.71 + 0.27x$. График је приказан на слици 2 и видимо да се слаже с подацима.

в) На основу узорка можемо израчунати $\bar{x} = 10$, $SSE = S_{yy} - \hat{\beta}S_{xy} = 0.544$ и $\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}} = 0.22$. За $x = 13$, заменом у једначину праве добијене под б), добијамо $\hat{\mu}_{Y|13} = 4.22$. Из таблице Студентове расподеле, за $\nu = n - 2 = 11$, добијамо $t_{\alpha/2} = 1.796$. Заменом у формулу за интервал поверења

$$I_{\mu_{Y|13}} = \left(\hat{\mu}_{Y|13} - t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(13-\bar{x})^2}{S_{xx}}}, \hat{\mu}_{Y|13} + t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(13-\bar{x})^2}{S_{xx}}}\right)$$
$$= \left(4.22 - 1.796 \cdot 0.22\sqrt{\frac{1}{13} + \frac{(13-10)^2}{262}}, 4.22 + 1.796 \cdot 0.22\sqrt{\frac{1}{13} + \frac{(13-10)^2}{262}}\right)$$

добијамо интервал (4.09,4.35).

г) Вредност од 60 дана не налази се у опсегу података (од 3 до 17) па не можемо проценити средњу дужину крила врапца од 60 дана.

(бонус) Линеарни модел нема смисла уколико је $\beta = 0$, тј. Y је исто за свако x, па бисмо стога тестирали хипотезу $H_0 : \beta = 0$ против алтернативе $H_1 : \beta \neq 0$.

5. [**12 поена**] Проучава се понашање једне врсте змија кад је угрожена од стране грабљивца. Пошто су змије хладнокрвне животиње, има индиција да фактор температуре има утицаја на понашање змије. Мери се растојање (у м) колико се близу може прићи змији пре него што онда побегне. Четрнаест змија подељено је у три групе и изложене су трима различитим температурама. Добијени су следећи резултати

температура		
23°	25°	27°
0.5	0.75	3.5
3.0	4.0	8.0
1.0	3.25	5.5
1.25	5.25	6.0
4.25	4.75	

Случајна променљива која представља растојање на које се може прићи змији нема нормалну расподелу.

а) [*2 поена*] Шта треба узети за нулту, а шта за алтернативну хипотезу?

б) [*7 поена*] С прагом значајности $\alpha = 0.05$ извршити одговарајуће тестирање користећи тест Крускал-Валиса.

в) [*3 поена*] Који се закључак може извести на основу резултата теста?

[*бонус*] Када би растојање имало нормалну расподелу, да ли бисте и тада могли користити тест Крускал-Валиса? Који бисте тест у том случају изабрали и зашто?

Решење.

а) Нека је средња вредност раздаљине на коју се може прићи змији при првој температури μ_1, при другој μ_2, а при трећој μ_3. Нулта хипотеза је да нема разлике у наведеним просечним вредностима ($H_0 : \mu_1 = \mu_2 = \mu_3$), док је алтернатива да су међу овим средњим вредностима бар две различите.

б) Како раздаљина на коју се може прићи змији нема нормалну расподелу, користимо тест Крускал-Валиса. Да бисмо одредили тест статистику, најпре све елементе узорка поређамо у ред и доделимо им рангове.

рангови елемената		
23°	25°	27°
1	2	7
5	8	14
3	6	12
4	11	13
9	10	
$R_1 = 22$	$R_2 = 37$	$R_3 = 46$

Тест статистика је $H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$. Из нашег узорка добијамо $h_0 = \frac{12}{14 \cdot 15}(\frac{22^2}{5} + \frac{37^2}{5} + \frac{46^2}{4}) - 3 \cdot 15 = 6.41$.

p-вредност теста рачунамо као површину испод графика χ_2 расподеле за $\nu = 2$ десно од h_0. Из таблице за χ^2 расподелу видимо да се 6.41 налази између 5,99148 и 7,37778, одакле закључујемо да је p-вредност теста између 0.025 и 0.05.

в) Како је p-вредност теста мања од 0.05, закључујемо да постоји разлика између средњих вредности за макар две температуре.

(бонус) У случају када би подаци имали нормалну расподелу, такође бисмо могли да користимо тест Крускал-Валиса. Међутим, у том случају бисмо радије користили једнофакторску дисперзиону анализу јер је то моћнија метода која успешније може открити разлике између средњих вредности одговарајућих група.