

Анализа података у биологији

Марија Цупарић

Први час



Предиспитне обавезе:

- ★ четири теста по 20 поена, рачунају се три најбоље урађена

Испит:

- ★ усмени испит (40 поена)

Литература:

- ★ Б. Милошевић. Основи статистике, Математички факултет, Београд, 2021.
- ★ J.S. Milton, J.J. Corbet and P.M. McTeer. Introduction to Statistics, DC Heath & Company, 1986.

- Материјали:**
- ★ <http://old.matf.bg.ac.rs/p/-marija-radicevic>
 - ★ ???



- ★ Статистика је наука о подацима.
- ★ Бави се њиховим прикупљањем, приказивањем, анализом и закључивањем на основу њих.
- ★ Задатак статистичара је да предложи одговарајући математички модел којим би се подаци адекватно описали.
- ★ Статистички модел је заправо скуп претпоставки, а једна од најзначајнијих јесте везана за расподелу обележја.



- ★ Статистика је наука о подацима.
- ★ Бави се њиховим прикупљањем, приказивањем, анализом и закључивањем на основу њих.
- ★ Задатак статистичара је да предложи одговарајући математички модел којим би се подаци адекватно описали.
- ★ Статистички модел је заправо скуп претпоставки, а једна од најзначајнијих јесте везана за расподелу обележја.

Увек смо заинтересовани за одређену групу објеката.
Желимо да опишемо неке карактеристике те групе или да донесемо неке закључке о њиховом понашању.



- ★ Статистика је наука о подацима.
- ★ Бави се њиховим прикупљањем, приказивањем, анализом и закључивањем на основу њих.
- ★ Задатак статистичара је да предложи одговарајући математички модел којим би се подаци адекватно описали.
- ★ Статистички модел је заправо скуп претпоставки, а једна од најзначајнијих јесте везана за расподелу обележја.

Увек смо заинтересовани за одређену групу објеката.

Желимо да опишемо неке карактеристике те групе или да донесемо неке закључке о њиховом понашању.

Међутим, група је обично велика да бисмо доносили закључке на основу ње целе.



- ★ Статистика је наука о подацима.
- ★ Бави се њиховим прикупљањем, приказивањем, анализом и закључивањем на основу њих.
- ★ Задатак статистичара је да предложи одговарајући математички модел којим би се подаци адекватно описали.
- ★ Статистички модел је заправо скуп претпоставки, а једна од најзначајнијих јесте везана за расподелу обележја.

Увек смо заинтересовани за одређену групу објеката.

Желимо да опишемо неке карактеристике те групе или да донесемо неке закључке о њиховом понашању.

Међутим, група је обично велика да бисмо доносили закључке на основу ње целе.

Због тога посматрамо само њен део.



популација (Ω): скуп објеката који се проучава у односу на неку променљиву квантитативну (број деце, оцена на испиту, тежина, висина,...) или квалитативну (крвна група, пол, боја,...) особину



популација (Ω): скуп објеката који се проучава у односу на неку променљиву квантитативну (број деце, оцена на испиту, тежина, висина,...) или квалитативну (крвна група, пол, боја,...) особину

Дефинисање циљне популације је важан и често тежак део студије. На пример, у политичкој анкети, да ли би циљна популација требало да буду

- ? сви одрасли људи са правом гласа,
- ? све особе које су гласале на прошлим изборима,
- ? или нешто треће,

зависи од тога шта нам је циљ.

популација (Ω): скуп објеката који се проучава у односу на неку променљиву квантитативну (број деце, оцена на испиту, тежина, висина,...) или квалитативну (крвна група, пол, боја,...) особину

Дефинисање циљне популације је важан и често тежак део студије. На пример, у политичкој анкети, да ли би циљна популација требало да буду

- ? сви одрасли људи са правом гласа,
- ? све особе које су гласале на прошлим изборима,
- ? или нешто треће,

зависи од тога шта нам је циљ.

обележје (X): особина или карактеристика која се проучава

Примери

(1) **Популација:** Бактерије у култури (нпр. у лабораторијским условима)

Обележја:

- ★ брзина раста популације
- ★ време дуплирања
- ★ отпорност на антибиотик
- ★ просечна величина ћелије
- ★ стопа мутације



Примери

- (1) **Популација:** Бактерије у култури (нпр. у лабораторијским условима)

Обележја:

- ★ брзина раста популације
- ★ време дуплирања
- ★ отпорност на антибиотик
- ★ просечна величина ћелије
- ★ стопа мутације

- (2) **Популација:** Пацијенти у клиничкој студији

Обележја:

- ★ концентрација биомаркера у крви
- ★ време опоравка
- ★ одговор на терапију
- ★ број нежељених реакција
- ★ ниво крвног притиска



Шта је заправо обележје?



Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.



Шта је заправо обележје?



Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.

★ означавамо је са $X, Y, Z, \dots$



Шта је заправо обележје?

Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.

- ★ означавамо је са $X, Y, Z, \dots$
- ★ скуп реализација случајне величине - скуп свих могућих вредности



Шта је заправо обележје?

Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.

- ★ означавамо је са $X, Y, Z, \dots$
- ★ скуп реализација случајне величине - скуп свих могућих вредности

Случајна променљива дефинисана на објектима популације назива се и **обележјем** те популације.



Шта је заправо обележје?

Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.

- ★ означавамо је са $X, Y, Z, \dots$
- ★ скуп реализација случајне величине - скуп свих могућих вредности

Случајна променљива дефинисана на објектима популације назива се и **обележјем** те популације.

Примери (наставак)

- (1) Различите бактерије могу имати различиту брзину раста због генетских разлика или услова средине, време дуплирања зависи од температуре и доступности хранљивих материја,...

Шта је заправо обележје?

Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.

- ★ означавамо је са $X, Y, Z, \dots$
- ★ скуп реализација случајне величине - скуп свих могућих вредности

Случајна променљива дефинисана на објектима популације назива се и **обележјем** те популације.

Примери (наставак)

- (1) Различите бактерије могу имати различиту брзину раста због генетских разлика или услова средине, време дуплирања зависи од температуре и доступности хранљивих материја,...
- (2) Код различитих пацијената концентрација биомаркера може варирати због разлика у метаболизму или стадијуму болести, време опоравка зависи од општег здравственог стања,...



Шта је заправо обележје?

Ако као резултат експеримента нека променљива може да узима различите вредности, које унапред нису познате, онда њу зовемо **случајна променљива** или **случајна величина**.

- ★ означавамо је са $X, Y, Z, \dots$
- ★ скуп реализација случајне величине - скуп свих могућих вредности

Случајна променљива дефинисана на објектима популације назива се и **обележјем** те популације.

Примери (наставак)

- (1) Различите бактерије могу имати различиту брзину раста због генетских разлика или услова средине, време дуплирања зависи од температуре и доступности хранљивих материја,...
- (2) Код различитих пацијената концентрација биомаркера може варирати због разлика у метаболизму или стадијуму болести, време опоравка зависи од општег здравственог стања,...



Пре сваког од експеримената из претходних примера не можемо са сигурношћу рећи који ће бити исход, након експеримента променљиве ће имати нумеричку вредност одређену насумично.



Случајне величине се деле на:

дискретне - могу узети највише коначно или пребројиво бесконачно много различитих вредности

апсолутно непрекидне - могу узети било коју вредност из неког интервала реалних бројева

оне које нису ниједно од претходна два



Случајне величине се деле на:

дискретне - могу узети највише коначно или пребројиво бесконачно много различитих вредности

апсолутно непрекидне - могу узети било коју вредност из неког интервала реалних бројева

оне које нису ниједно од претходна два

Дискретне случајне величине могу бити:

§ категоричке (квалитативне):

- ★ номиналне (крвна група, пол, сексуално опредељење,...)
- ★ ординалне (платни разред, степен стручне спреме,...)

§ нумеричке (квантитативне)



Случајне величине се деле на:

дискретне - могу узети највише коначно или пребројиво бесконачно много различитих вредности

апсолутно непрекидне - могу узети било коју вредност из неког интервала реалних бројева

оне које нису ниједно од претходна два

Дискретне случајне величине могу бити:

§ категоричке (квалитативне):

- ★ номиналне (крвна група, пол, сексуално опредељење,...)
- ★ ординалне (платни разред, степен стручне спреме,...)

§ нумеричке (квантитативне)

Непрекидне променљиве све припадају групи нумеричких.



Случајне величине се деле на:

дискретне - могу узети највише коначно или пребројиво бесконачно много различитих вредности

апсолутно непрекидне - могу узети било коју вредност из неког интервала реалних бројева

оне које нису ниједно од претходна два

Дискретне случајне величине могу бити:

§ категоричке (квалитативне):

- ★ номиналне (крвна група, пол, сексуално опредељење,...)
- ★ ординалне (платни разред, степен стручне спреме,...)

§ нумеричке (квантитативне)

Непрекидне променљиве све припадају групи нумеричких.

Иако се категоричка обележја често представљају нумерички („кодирају се”), анализа категоричких и нумеричких обележја се не врши на исти начин. Разлика између номиналног и ординалног типа је та што су у првом категорије равноправне док код другог имају поредак.



Примери (наставак) Типови случајних величина

- (1)
 - ★ брзина раста популације - непрекидна
 - ★ време дуплирања - непрекидна
 - ★ отпорност на антибиотик - дискретна (категоричка номинална, нпр. „осетљива“ / „умерено отпорна“ / „отпорна“)
 - ★ просечна величина ћелије - непрекидна
 - ★ стопа мутације - непрекидна



Примери (наставак) Типови случајних величина

- (1)
 - ★ брзина раста популације - непрекидна
 - ★ време дуплирања - непрекидна
 - ★ отпорност на антибиотик - дискретна (категоричка номинална, нпр. „осетљива“ / „умерено отпорна“ / „отпорна“)
 - ★ просечна величина ћелије - непрекидна
 - ★ стопа мутације - непрекидна
- (2)
 - ★ концентрација биомаркера у крви - непрекидна
 - ★ време опоравка - непрекидна
 - ★ број нежељених реакција - дискретна (нумеричка)
 - ★ ниво крвног притиска - непрекидна
 - ★ одговор на терапију - дискретна (категоричка номинална са категоријама нпр. „без одговора“ / „делимичан одговор“ / „потпун одговор“)



Узорак $(X_1, \dots, X_n)$: подскуп популације на основу којег доносимо закључке, представља вредности обележја на изабраним јединкама



Узорак $(X_1, \dots, X_n)$: подскуп популације на основу којег доносимо закључке, представља вредности обележја на изабраним јединкама

★ случајан - ако се подскуп бира насумично



Узорак $(X_1, \dots, X_n)$: подскуп популације на основу којег доносимо закључке, представља вредности обележја на изабраним јединкама

- ★ случајан - ако се подскуп бира насумично
- ★ репрезентативан - добро осликава популацију, на основу њега можемо да доносимо закључке о читавој популацији



Узорак $(X_1, \dots, X_n)$: подскуп популације на основу којег доносимо закључке, представља вредности обележја на изабраним јединкама

- ★ случајан - ако се подскуп бира насумично
- ★ репрезентативан - добро осликава популацију, на основу њега можемо да доносимо закључке о читавој популацији
- ★ прост случајан узорак обима n - n -димензиона случајна величина $(X_1, \dots, X_n)$ при чему су компоненте независне и све имају исту расподелу као обележје



Узорак $(X_1, \dots, X_n)$: подскуп популације на основу којег доносимо закључке, представља вредности обележја на изабраним јединкама

- ★ случајан - ако се подскуп бира насумично
- ★ репрезентативан - добро осликава популацију, на основу њега можемо да доносимо закључке о читавој популацији
- ★ прост случајан узорак обима n - n -димензиона случајна величина $(X_1, \dots, X_n)$ при чему су компоненте независне и све имају исту расподелу као обележје

Касније током курса ћемо се упознати са појмовима независности и расподеле обележја.

Узорак $(X_1, \dots, X_n)$: подскуп популације на основу којег доносимо закључке, представља вредности обележја на изабраним јединкама

- ★ случајан - ако се подскуп бира насумично
- ★ репрезентативан - добро осликава популацију, на основу њега можемо да доносимо закључке о читавој популацији
- ★ прост случајан узорак обима n - n -димензиона случајна величина $(X_1, \dots, X_n)$ при чему су компоненте независне и све имају исту расподелу као обележје

Касније током курса ћемо се упознати са појмовима независности и расподеле обележја.

- ★ реализовани узорак $(x_1, \dots, x_n)$ - конкретан низ вредности обележја добијених на елементима популације који су узети у узорак

Параметар популације (θ): нека описна мера случајне величине (обележја) посматране на целој популацији



Параметар популације (θ): нека описна мера случајне величине (обележја) посматране на целој популацији

Примери (наставак) Параметри популације

- (1) μ просечна брзина раста популације, μ_T просечно време дуплирања, p_i вероватноћа i -те категорије отпорности на антибиотик
- (2) μ просечна концентрација биомаркера у крви, λ просечан број нежељених реакција, p_i вероватноћа i -те категорије одговора на терапију ✓

Параметар популације (θ): нека описна мера случајне величине (обележја) посматране на целој популацији

Примери (наставак) Параметри популације

- (1) μ просечна брзина раста популације, μ_T просечно време дуплирања, p_i вероватноћа i -те категорије отпорности на антибиотик
- (2) μ просечна концентрација биомаркера у крви, λ просечан број нежељених реакција, p_i вероватноћа i -те категорије одговора на терапију ✓

Статистика ($T_n = g(X_1, \dots, X_n)$): описна мера случајне величине (обележја) посматране само на узорку

Статистика зависи од узорка и константи, а не сме зависити од непознатих параметара.

Кораци у статистичкој анализи



- (1) осмишљавање експеримента
 - ▷ одредити популацију која се проучава
 - ▷ поставити питања у вези популације на која желимо одговор
 - ▷ одредити случајне величине (обележја) чије ће проучавање помоћи да дођемо до одговора
 - ▷ одредити параметре популације који су од важности
- (2) узорковање и прикупљање података
- (3) прелиминарна анализа
- (4) идентификација аутлајера
- (5) конструкција статистичког модела
- (6) одговор на питања постављена у кораку (1)



Кораци у статистичкој анализи

- (1) осмишљавање експеримента
 - ▷ одредити популацију која се проучава
 - ▷ поставити питања у вези популације на која желимо одговор
 - ▷ одредити случајне величине (обележја) чије ће проучавање помоћи да дођемо до одговора
 - ▷ одредити параметре популације који су од важности
- (2) узорковање и прикупљање података
- (3) прелиминарна анализа
- (4) идентификација аутлајера
- (5) конструкција статистичког модела
- (6) одговор на питања постављена у кораку (1)

Циљ

- ★ пронаћи што једноставнији статистички модел који довољно добро описује популацију
- ★ оправдати коришћење модела



ПРЕЛИМИНАРНА АНАЛИЗА

- ★ има улогу да опише податке на једноставан начин



ПРЕЛИМИНАРНА АНАЛИЗА

- ★ има улогу да опише податке на једноставан начин
- ★ њен саставни део су графички приказ података и приказ сумарних статистика које нам могу дати више информација о нумеричким карактеристикама обележја



ПРЕЛИМИНАРНА АНАЛИЗА

- ★ има улогу да опише податке на једноставан начин
- ★ њен саставни део су графички приказ података и приказ сумарних статистика које нам могу дати више информација о нумеричким карактеристикама обележја
- ★ једноставније речено у овом кораку користимо **дескриптивне статистике**



- ★ има улогу да опише податке на једноставан начин
- ★ њен саставни део су графички приказ података и приказ сумарних статистика које нам могу дати више информација о нумеричким карактеристикама обележја
- ★ једноставније речено у овом кораку користимо **дескриптивне статистике**

За приказ узорка у случају категоричког обележја можемо користити

табеларни приказ - приказује учесталост по категоријама

тракасти дијаграм (barplot) - фреквенције приказане у табели се приказују у виду трака (стубова) на графику

кружни дијаграм (pie chart) - круг се дели на исечке коју су пропорционални фреквенцијама приказаним у табели



База попис





```
> baza$radni.st<-factor(baza$radni.st)
> baza$bracno.st<-factor(baza$bracno.st)
> baza$br.dece<-as.numeric(baza$br.dece)
> baza$godine<-as.numeric(baza$godine)
> baza$obrazov<-factor(baza$obrazov, ordered = TRUE)
> baza$otac.obr<-factor(baza$otac.obr, ordered = TRUE)
> baza$majka.obr<-factor(baza$majka.obr, ordered = TRUE)
> baza$pol<-factor(baza$pol)
> summary(baza)
```

	radni.st	bracno.st	br.dece	godine	obrazov	otac.obr	majka.obr	pol
1	:237	1:232	Min. : 1.000	Min. : 1.00	0: 52	0 :147	1 :179	1:185
2	: 58	2: 31	1st Qu.: 1.000	1st Qu.:16.00	1:208	1 :135	0 :152	2:215
5	: 40	3: 52	Median : 3.000	Median :25.00	2: 29	2 : 5	3 : 24	
7	: 29	4: 6	Mean : 2.868	Mean :27.14	3: 74	3 : 36	8 : 16	
4	: 17	5: 79	3rd Qu.: 4.000	3rd Qu.:36.00	4: 37	4 : 16	2 : 10	
3	: 9		Max. :10.000	Max. :65.00		8 : 13	4 : 9	
(other): 10						NA: 48	(other): 10	





```
> baza$radni.st<-factor(baza$radni.st)
> baza$bracno.st<-factor(baza$bracno.st)
> baza$br.dece<-as.numeric(baza$br.dece)
> baza$godine<-as.numeric(baza$godine)
> baza$obrazov<-factor(baza$obrazov, ordered = TRUE)
> baza$otac.obr<-factor(baza$otac.obr, ordered = TRUE)
> baza$majka.obr<-factor(baza$majka.obr, ordered = TRUE)
> baza$pol<-factor(baza$pol)
> summary(baza)
```

	radni.st	bracno.st	br.dece	godine	obrazov	otac.obr	majka.obr	pol
1	:237	1:232	Min. : 1.000	Min. : 1.00	0: 52	0 :147	1 :179	1:185
2	: 58	2: 31	1st Qu.: 1.000	1st Qu.:16.00	1:208	1 :135	0 :152	2:215
5	: 40	3: 52	Median : 3.000	Median :25.00	2: 29	2 : 5	3 : 24	
7	: 29	4: 6	Mean : 2.868	Mean :27.14	3: 74	3 : 36	8 : 16	
4	: 17	5: 79	3rd Qu.: 4.000	3rd Qu.:36.00	4: 37	4 : 16	2 : 10	
3	: 9		Max. :10.000	Max. :65.00		8 : 13	4 : 9	
(other): 10						NA: 48	(other): 10	

Пре даље анализе требало би избацити недостајуће податке



```
> baza1<-na.omit(baza)
> summary(baza1)
```

За табелирање једне променљиве користи се



```
> levels(baza1$radni.st)=c("puno radno vreme", "skraceno radno  
vreme", "trenutno ne radi", "nezaposlen, otpusten", "u penziji",  
"ucenik/ca", "domacin/ca", "drugo")  
> table(baza1$radni.st)
```



За табелирање једне променљиве користи се



```
> levels(baza1$radni.st)=c("puno radno vreme", "skraceno radno  
vreme", "trenutno ne radi", "nezaposlen, otpusten", "u penziji",  
"ucenik/ca", "domacin/ca", "drugo")  
> table(baza1$radni.st)
```

puno radno vreme	skraceno radno vreme	trenutno ne radi	nezaposlen, otpusten	u penziji
205	56	7	15	31
ucenik/ca	domacin/ca	drugo		
6	23	3		



За табелирање једне променљиве користи се



```
> levels(baza1$radni.st)=c("puno radno vreme", "skraceno radno  
vreme", "trenutno ne radi", "nezaposlen, otpusten", "u penziji",  
"ucenik/ca", "domacin/ca", "drugo")  
> table(baza1$radni.st)
```

puno radno vreme	skraceno radno vreme	trenutno ne radi	nezaposlen, otpusten	u penziji
205	56	7	15	31
ucenik/ca	domacin/ca	drugo		
6	23	3		

или једне променљиве у односу на другу



```
> levels(baza1$pol)=c("M","Z")  
> table(baza1$radni.st, baza$pol)
```



За табелирање једне променљиве користи се

R

```
> levels(baza1$radni.st)=c("puno radno vreme", "skraceno radno  
vreme", "trenutno ne radi", "nezaposlen, otpusten", "u penziji",  
"ucenik/ca", "domacin/ca", "drugo")  
> table(baza1$radni.st)
```

puno radno vreme	skraceno radno vreme	trenutno ne radi	nezaposlen, otpusten	u penziji
205	56	7	15	31
ucenik/ca	domacin/ca	drugo		
6	23	3		

или једне променљиве у односу на другу

R

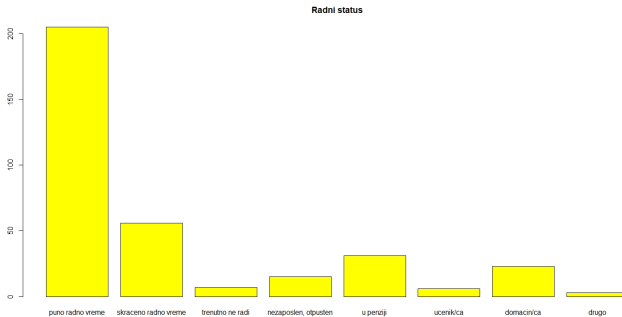
```
> levels(baza1$pol)=c("M", "Z")  
> table(baza1$radni.st, baza$pol)
```

	M	Z
puno radno vreme	118	87
skraceno radno vreme	14	42
trenutno ne radi	1	6
nezaposlen, otpusten	9	6
u penziji	15	16
ucenik/ca	1	5
domacin/ca	0	23
drugo	3	0

Тракасти дијаграм на x -оси приказује категорије, а на y -оси фраквенције појављивања елемената из узорка у свакој од категорија

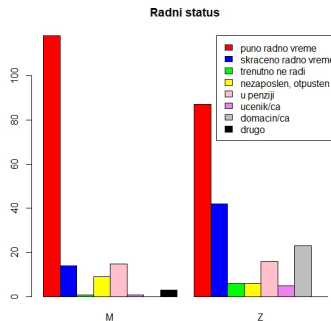
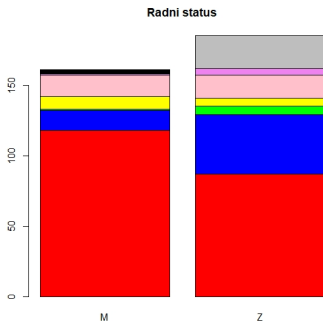
R

```
> barplot(table(baza1$radni.st), col="yellow", main="Radni status")
```



Може се цртати једна променљива у односу на категорије друге променљиве.

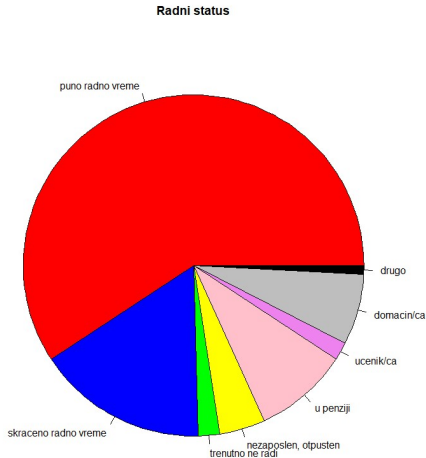
```
> boje<-c("red","blue","green","yellow","pink","violet","grey","black")
> barplot(table(baza1$radni.st, baza1$pol), main="Radni status", col=boje)
> barplot(table(baza1$radni.st, baza1$pol), main="Radni status", col=boje,
  beside=TRUE)
> legend("topright", legend=levels(baza1$radni.st), fill=boje,cex=0.8)
```



Кружни дијаграм

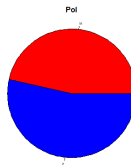
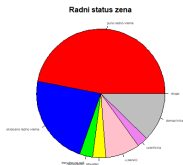
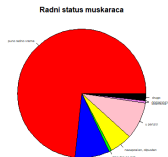
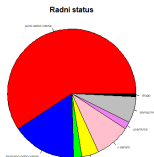
R

```
> pie(table(baza1$radni.st), main="Radni status", col=boje)
```



R

```
> par(mfrow=c(2,2))
> pie(table(baza1$radni.st), main="Radni status", col=boje, radius=1,
    cex=0.5)
> pie(table(baza1$radni.st[baza1$pol=="M"]), main="Radni status
    muskaraca", col=boje, radius=1, cex=0.5)
> pie(table(baza1$radni.st[baza1$pol=="F"]), main="Radni status
    zena", col=boje, radius=1, cex=0.5)
> pie(table(baza1$pol), main="Pol", col=boje, radius=1, cex=0.5)
```



ПРЕЛИМИНАРНА АНАЛИЗА НУМЕРИЧКИХ ОБЕЛЕЖЈА:

- ★ хистограм
- ★ дескриптивне статистике
 - ★ мере централне тенденције (мере положаја)
 - ★ мере расејања

Занима нас:

- ? Какав је облик расподеле? Да ли вредности случајне променљиве чине неку препознатљиву структуру? Који је положај података, тј. око које централне вредности су они распоређени?
- ? Колико има одступања међу подацима? Да ли су они прилично расејани или згуснути око централне вредности?



Хистограм

Гrafички приказ учесталости елемената узорка по категоријама.



Графички приказ учесталости елемената узорка по категоријама.

Поступак цртања

- ★ податке груписати у категорије (интервале), тако да сваки елемент из узорка припада тачно једној категорији



Графички приказ учесталости елемената узорка по категоријама.

Поступак цртања

- ★ податке груписати у категорије (интервале), тако да сваки елемент из узорка припада тачно једној категорији
 - ▷ препорука је да буде бар 5 категорија и да је број категорија $\lceil \log_2 n \rceil + 1$, где је n обим узорка
 - ▷ за леву границу првог интервала узети вредност мало мања од минималног елемента узорка
 - ▷ стварна ширина стуба је заокружену вредност (на већу вредност и са истим бројем децимала као подаци) количника између распона (разлика између минималне и максималне вредности у узорку) и броја категорија



Графички приказ учесталости елемената узорка по категоријама.

Поступак цртања

- ★ податке груписати у категорије (интервале), тако да сваки елемент из узорка припада тачно једној категорији
 - ▷ препорука је да буде бар 5 категорија и да је број категорија $\lceil \log_2 n \rceil + 1$, где је n обим узорка
 - ▷ за леву границу првог интервала узети вредност мало мања од минималног елемента узорка
 - ▷ стварна ширина стуба је заокружену вредност (на већу вредност и са истим бројем децимала као подаци) количника између распона (разлика између минималне и максималне вредности у узорку) и броја категорија
- ★ одредити број елемената из узорка који се налазе у свакој категорији



Графички приказ учесталости елемената узорка по категоријама.

Поступак цртања

- ★ податке груписати у категорије (интервале), тако да сваки елемент из узорка припада тачно једној категорији
 - ▷ препорука је да буде бар 5 категорија и да је број категорија $\lceil \log_2 n \rceil + 1$, где је n обим узорка
 - ▷ за леву границу првог интервала узети вредност мало мања од минималног елемента узорка
 - ▷ стварна ширина стуба је заокружену вредност (на већу вредност и са истим бројем децимала као подаци) количника између распона (разлика између минималне и максималне вредности у узорку) и броја категорија
- ★ одредити број елемената из узорка који се налазе у свакој категорији
- ★ уцртати положај категорија



Графички приказ учесталости елемената узорка по категоријама.

Поступак цртања

- ★ податке груписати у категорије (интервале), тако да сваки елемент из узорка припада тачно једној категорији
 - ▷ препорука је да буде бар 5 категорија и да је број категорија $\lceil \log_2 n \rceil + 1$, где је n обим узорка
 - ▷ за леву границу првог интервала узети вредност мало мања од минималног елемента узорка
 - ▷ стварна ширина стуба је заокружену вредност (на већу вредност и са истим бројем децимала као подаци) количника између распона (разлика између минималне и максималне вредности у узорку) и броја категорија
- ★ одредити број елемената из узорка који се налазе у свакој категорији
- ★ уцртати положај категорија
- ★ нацртати „правоугаонике” у одговарајућем облику

Врсте хистограма:

хистограм апсолутних фреквенција: на y —оси су фреквенције, односно висина i —тог правоугаоника је број елемената из узорка који се налазе у i —тој категорији, n_i ;



Врсте хистограма:

хистограм апсолутних фреквенција: на y —оси су фреквенције, односно висина i —тог правоугаоника је број елемената из узорка који се налазе у i —тој категорији, n_i ;

хистограм релативних фреквенција: на y —оси су фреквенције подељене величином узорка, односно висина i —тог правоугаоника је $\frac{n_i}{n}$;



Врсте хистограма:

хистограм апсолутних фреквенција: на y —оси су фреквенције, односно висина i —тог правоугаоника је број елемената из узорка који се налазе у i —тој категорији, n_i ;

хистограм релативних фреквенција: на y —оси су фреквенције подељене величином узорка, односно висина i —тог правоугаоника је $\frac{n_i}{n}$;

хистограм густине: висина i —тог правоугаоника је $\frac{n_i}{nd_i}$, где је d_i дужина i —тог интервала.



Врсте хистограма:

хистограм апсолутних фреквенција: на y —оси су фреквенције, односно висина i —тог правоугаоника је број елемената из узорка који се налазе у i —тој категорији, n_i ;

хистограм релативних фреквенција: на y —оси су фреквенције подељене величином узорка, односно висина i —тог правоугаоника је $\frac{n_i}{n}$;

хистограм густине: висина i —тог правоугаоника је $\frac{n_i}{nd_i}$, где је d_i дужина i —тог интервала.

Проблеми хистограма:

- ★ његов изглед зависи од положаја и броја категорија,
- ★ функција која се црта је део-по-део константна а не непрекидна.



Врсте хистограма:

хистограм апсолутних фреквенција: на y —оси су фреквенције, односно висина i —тог правоугаоника је број елемената из узорка који се налазе у i —тој категорији, n_i ;

хистограм релативних фреквенција: на y —оси су фреквенције подељене величином узорка, односно висина i —тог правоугаоника је $\frac{n_i}{n}$;

хистограм густине: висина i —тог правоугаоника је $\frac{n_i}{nd_i}$, где је d_i дужина i —тог интервала.

Проблеми хистограма:

- ★ његов изглед зависи од положаја и броја категорија,
- ★ функција која се црта је део-по-део константна а не непрекидна.

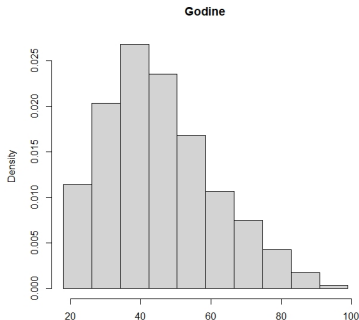
Хистограм густине заправо представља оцену густине обележја за које претпостављамо да има апсолутно непрекидну расподелу, па се идејно може видети које расподеле долазе у обзир за посматране податке



Одређивање граница хистограма по формули

R

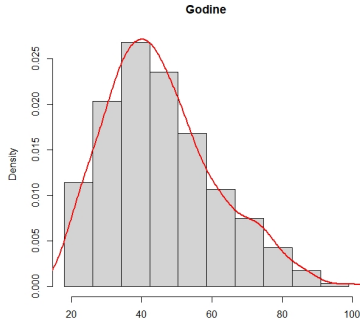
```
> broj.kategorija <- ceiling(log(length(baza1$godine),2))+1
> d<-(max(baza1$godine)-min(baza1$godine))/broj.kategorija
> granice<-min(baza1$godine)-1+(0:broj.kategorija)*(d+0.1)
> hist(baza1$godine, breaks=granice, prob=TRUE)
```



Додавање оцене густине на хистрограм

R

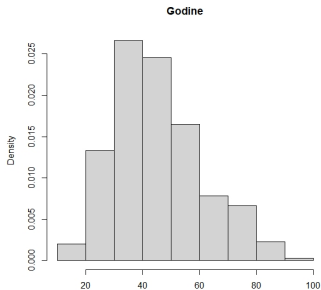
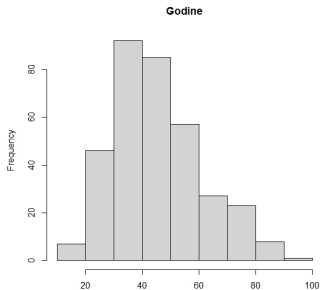
```
> broj.kategorija <- ceiling(log(length(baza1$godine),2))+1
> d<-(max(baza1$godine)-min(baza1$godine))/broj.kategorija
> granice<-min(baza1$godine)-1+(0:broj.kategorija)*(d+0.1)
> hist(baza1$godine, breaks=granice, prob=TRUE)
> lines(density(baza1$godine), col="red", lwd=2)
```



Коришћење подразумеваних граница функције hist

R

```
> hist(baza1$godine, main="Godine", xlab="")
> hist(baza1$godine, main="Godine", prob=TRUE, xlab="")
```



Процена удела популације чија је старост мања од 40 година

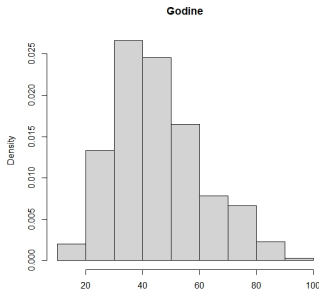
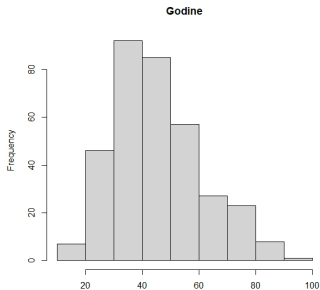
R

```
> H<-hist(baza1$godine, prob=TRUE, plot=FALSE)
> cumsum(H$density*diff(H$breaks))
[1] 0.02023121 0.15317919 0.41907514 0.66473988 0.82947977 0.90751445
0.97398844 0.99710983 1.00000000
```

Коришћење подразумеваних граница функције hist

R

```
> hist(baza1$godine, main="Godine", xlab="")  
> hist(baza1$godine, main="Godine", prob=TRUE, xlab="")
```



Процена удела популације чија је старост мања од 40 година

R

```
> H<-hist(baza1$godine, prob=TRUE, plot=FALSE)  
> cumsum(H$density*diff(H$breaks))  
[1] 0.02023121 0.15317919 0.41907514 0.66473988 0.82947977 0.90751445  
0.97398844 0.99710983 1.00000000
```

Тражена процена је 0.42.

О положају расподеле се може сазнати и из следећих параметара

- ★ средња вредност популације
- ★ медијана популације
- ★ мода популације



О положају расподеле се може сазнати и из следећих параметара

- ★ средња вредност популације
- ★ медијана популације
- ★ мода популације

Ови параметри су непознати и сваки од њих процењујемо одговарајућом статистиком.



Средња вредност

- ★ Други називи: очекивана вредност, математичко очекивање
- ★ Најчешћа ознака μ
- ★ Процењујемо га узорачком средином $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$



- ★ Други називи: очекивана вредност, математичко очекивање
- ★ Најчешћа ознака μ
- ★ Процењујемо га узорачком средином $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$

R `mean(x)`



- ★ Други називи: очекивана вредност, математичко очекивање
- ★ Најчешћа ознака μ
- ★ Процењујемо га узорачком средином $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$

R `mean(x)`

Пример. Број упамћених речи за два минута: 8 2 4 9 7 2 12 5 5 7

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{61}{10} = 6.1.$$



Комбиновање више узорачких средина

Пример. Број хитних случајева у једној болници је $\bar{x}_1 = 3$ за $n_1 = 5$, а у другој болници $\bar{x}_2 = 15$ за $n_2 = 100$.

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{5 \cdot 3 + 100 \cdot 15}{5 + 100} = 14.4$$



```
> sredine <- c(3,15)
> tezine <- c(5,100)/105
> weighted.mean(sredine, tezine)
```



Медијана

- ★ Медијана је средишња тачка, односно вредност од које је пола популације веће, а пола мање



Медијана

- ★ Медијана је средишња тачка, односно вредност од које је пола популације веће, а пола мање
- ★ Ознака m_e



- ★ Медијана је средишња тачка, односно вредност од које је пола популације веће, а пола мање
- ★ Ознака m_e
- ★ Узорачка медијана је она тачка која се налази на средини узорка, односно, за варијациони низ $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, то је

$$\hat{m}_e = \begin{cases} X_{(k+1)}, & n = 2k + 1, \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k. \end{cases}$$



- ★ Медијана је средишња тачка, односно вредност од које је пола популације веће, а пола мање
- ★ Ознака m_e
- ★ Узорачка медијана је она тачка која се налази на средини узорка, односно, за варијациони низ $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, то је

$$\hat{m}_e = \begin{cases} X_{(k+1)}, & n = 2k + 1, \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k. \end{cases}$$



`median(x)`



- ★ Медијана је средишња тачка, односно вредност од које је пола популације веће, а пола мање
- ★ Ознака m_e
- ★ Узорачка медијана је она тачка која се налази на средини узорка, односно, за варијациони низ $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, то је

$$\hat{m}_e = \begin{cases} X_{(k+1)}, & n = 2k + 1, \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k. \end{cases}$$

 `median(x)`

Пример. Године старости купаца у једној продавници гардеробе

Жене	12	15	17	20	24	27	30	35	42	60
Мушкарци	17	29	37	40	72					

Медијана старости жена је $\frac{24+27}{2} = 25.5$, а мушкараца је 37 година.

Пример. Узорак тржишне вредности (у хиљадама доларима) десет кућа у једном насељу: 82 91 78.5 86 80.5 85 82.5 80 77 850.
Какав је ово крај?



Пример. Узорак тржишне вредности (у хиљадама доларима) десет кућа у једном насељу: 82 91 78.5 86 80.5 85 82.5 80 77 850.
Какав је ово крај?

Средња вредност је 159.25, а медијана је 82.25.



Пример. Узорак тржишне вредности (у хиљадама доларима) десет кућа у једном насељу: 82 91 78.5 86 80.5 85 82.5 80 77 850.
Какав је ово крај?

Средња вредност је 159.25, а медијана је 82.25.

Из вредности $\bar{x}$ извлачимо погрешан закључак о вредности кућа у крају, медијана нам даје много бољу информацију. То је због утицаја неуобичајене вредности 850 коју називамо аутлајером (енгл. outlier - онај који се ту налази али не припада).



- ★ Мода је најчешћа вредност у популацији



- ★ Мода је најчешћа вредност у популацији
- ★ Ако постоји више (локалних) максимума, кажемо да је расподела више модална па самим тим ова величина није јединствена.



- ★ Мода је најчешћа вредност у популацији
- ★ Ако постоји више (локалних) максимума, кажемо да је расподела више модална па самим тим ова величина није јединствена.
- ★ Процењујемо је узорачком модом која је најфреквентнији елемент у узорку.



- ★ Мода је најчешћа вредност у популацији
- ★ Ако постоји више (локалних) максимума, кажемо да је расподела више модална па самим тим ова величина није јединствена.
- ★ Процењујемо је узорачком модом која је најфреквентнији елемент у узорку.
- ★ У случају апсолутно непрекидних расподела све вредности су различине (свима је фреквенција 1) па ова оцена није добра.



- ★ Мода је најчешћа вредност у популацији
- ★ Ако постоји више (локалних) максимума, кажемо да је расподела више модална па самим тим ова величина није јединствена.
- ★ Процењујемо је узорачком модом која је најфреквентнији елемент у узорку.
- ★ У случају апсолутно непрекидних расподела све вредности су различите (свима је фреквенција 1) па ова оцена није добра.
- ★ Ова оцена нема уграђену функцију у R —у.



- ★ Мода је најчешћа вредност у популацији
- ★ Ако постоји више (локалних) максимума, кажемо да је расподела више модална па самим тим ова величина није јединствена.
- ★ Процењујемо је узорачком модом која је најфреквентнији елемент у узорку.
- ★ У случају апсолутно непрекидних расподела све вредности су различине (свима је фреквенција 1) па ова оцена није добра.
- ★ Ова оцена нема уграђену функцију у R —у.

Ако је расподела симетрична, медијана и очекивана вредност се поклапају. Ако је расподела додатно и унимодална, онда се све три мере централне тенденције поклапају. Одговарајуће статистике, наравно неће се поклапати, али ће имати блиске вредности.

