Статистички софтвер 4

Трећи час

Марија Радичевић

Математички факултет, Београд

2015.

Анализа података обухвата низ техника и метода, које служе као моћни алат за добијање значајних информација из података који су добијени било из примарних или секундарних извора података.

Комплетан истраживачки пројекат укључује:

- припремање прелиминарног плана анализе података
- преглед упитника и валидацију података
- едитовање (кориговање)
- кодирање
- пречишћавање и табелирање података

Припремање података за анализу

Квалитет улазних података одређује квалитет информација које се добијају у поступку анализе података, а на основу којих се доносе одговарајуће пословне одлуке.

Поступак припремања података за анализу:

1. Прелиминарни план анализе података
2. Преглед упитника и валидација података
3. Уређивање података
4. Кодирање података
5. Унос података
6. Пречишћавање и статистичко модификовање података
7. Табелирање података
8. Селекција стратегије анализа података

Преглед упитника и валидација података

Провера прихватљивости упитника

- непотпуни делови упиника
- системски или случајно прескочена питања
- неадекватност упитника у смислу задатког типа узорка
- мала варијанса одговора

Валидација података подразумева проверавање опсервација, упиника или целокупног испитивања. Обухвата:

1. скрининг
2. процедура
3. потпуност
4. превара

Уређивање (едитовање) података

Проверавање података у смислу постојања грешака, било да се ради о грешкама испитивача или испитаника. Проблеми који се могу идентификовати током уређивања података:

- грешке испитивача
- испитаник није одговорио на поједина питања
- нејасан одговор испитаника
- логички неконзистентни одговори
- невољност испитаника да одговори на питања
- неадеквантан испитаник

Третирање незадовољавајућих вредности

1. поновити контакт на терену
2. додељивање третмана недостајућих вредности
3. искључивање испитаника

Шифровање (кодирање) података

Додељивање одређене шифре сваком могућем одговору на постављено питање. Услови који морају бити испуњени:

услов искључености категорије одговора морају бити међусобно искључиве

услов потпуности сваки одговор на постављено питање може да се додели одређеној категорији

Унос података

- електронски облик
- остварује се преносом кодираних података из упитника у рачунарски систем
- посебно писани документи - шифарници
- формат радних табела (*Excel*, разни софтверски пакети)

Пречишћавање и статистичка модификација података

Пречишћавање података се односи на проверу конзистентности података и третман недостајућих вредности.
Провера конзистентности:

- логичка неконзистентност
- вредности ван дозвољеног интервала
- екстремне вредности

Третман недостајућих вредности:

- замена недостајућих вредности неком неутралном вредношћу
- брисање опсервација
- брисање података само за појединачне променљиве у оквиру посматране опсервације

Пречишћавање и статистичка модификација података

Статистичко модификовање података може се вршити на различите начине, а најчеши су:

- пондерисање података
- прекодирање података
- трансформација података мерених на различитим скалама мерења

Табелирање података

Процес рачунања броја опсервација које припадају одређеној категорији посматраног обележја (промењљиве).

- једноструко табелирање
- унакрсно табелирање

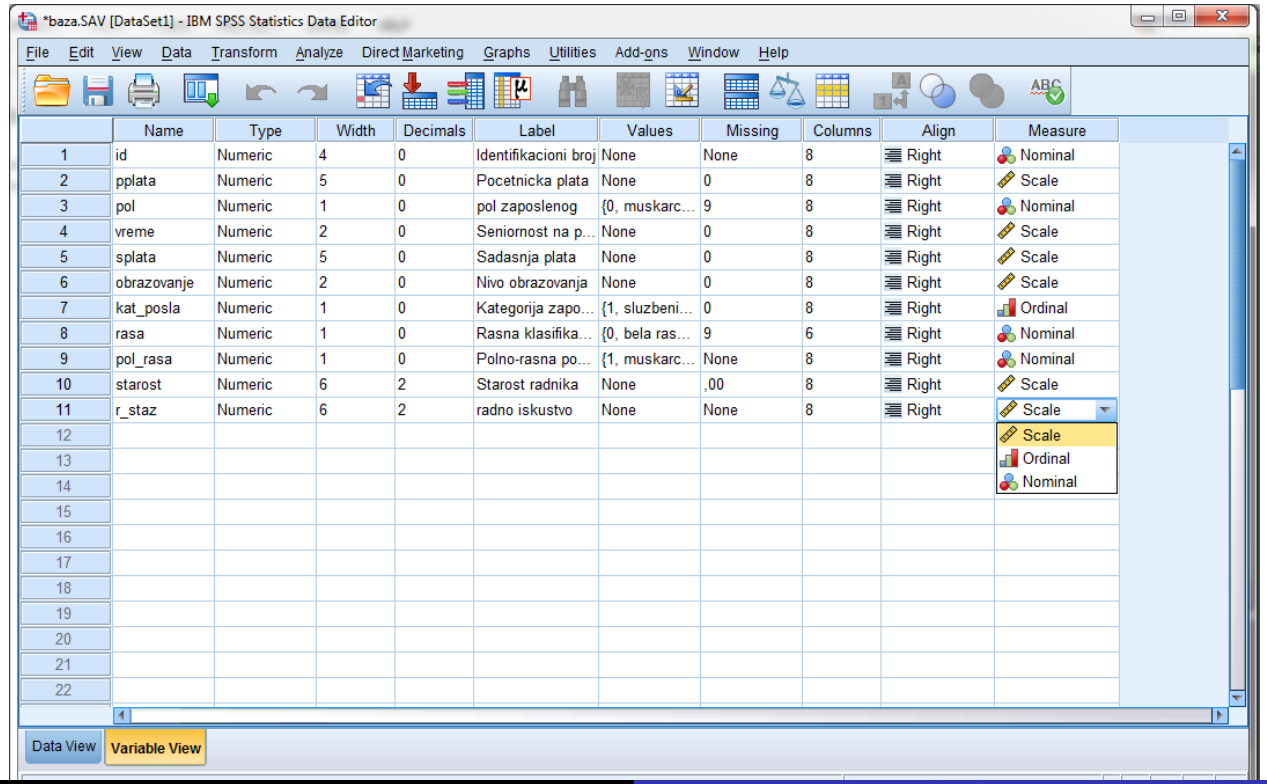
Избор стратегије анализе података

Фактори који утичу на избор стратегије анализе података:

- дефинисање проблема истраживања, приступ истраживању и дизајн истраживања
  - међусобна (не)зависност опсервација
  - број група које се посматрају
  - број мерења (промењљивих) по једном објекту посматрања у истраживању
  - могућност контроле посматраних промењљивих
- карактеристике података
- особине статистичких техника
- приступ истраживача

Карактеристике података - мерне скале

- номинална --→ Nominal
- ординална --→ Ordinal
- интервална --→ Scale
- релациона --→ Scale



Номинална скала

- симбол означава припадност објекта или лица одређеној групи (категорији)
- није уведена релација поретка (нерангиране категорије)
- дефинисана је само релација једнакости
- нумеричког или алфанумеричког типа у *SPSS*−у

Примери:

пол (0 =мушки, 1 =женски), **брачни статус** (1 =ожењен/удата, 2 =неожењен/неудата, 3 =разведен/разведена, 4 =удовац/удовица, 5 =остало), **боја очију, поштански број,...**

Ординална скала

- категоричка променљива са дефинисаном релацијом поретка (рангиране категорије)
- јединица мере није дефинисана (нпр. 3 је веће од 2, али се не зна интензитет разлике)
- интервали између узастопних вредности не морају бити исти (бројеви означавају само рангове)
- препоручује се да подаци буду нумеричког типа

Примери:

оцене (1 =недовољан, 2 =довољан, 3 =добар, 4 =врло добар, 5 =одличан), **позиција на ранг листи** (број бодова између позиција не мора бити исти),...

Интервална скала

- постоји поредак
- интервали између узастопних вредности су једнаки
- нула нема природно значење, тј. не представља одсуство посматраног својства
- има смисла посматрати разлику вредности, али не и колочник
- увек нумеричког типа

Примери:

стандардизовани *IQ* тестови, **температура ваздуха** (нпр., $12°$ је топлије од $4°$ за $8°$, али није три пута топлије јер $0°$ не представља одсуство температуре),...

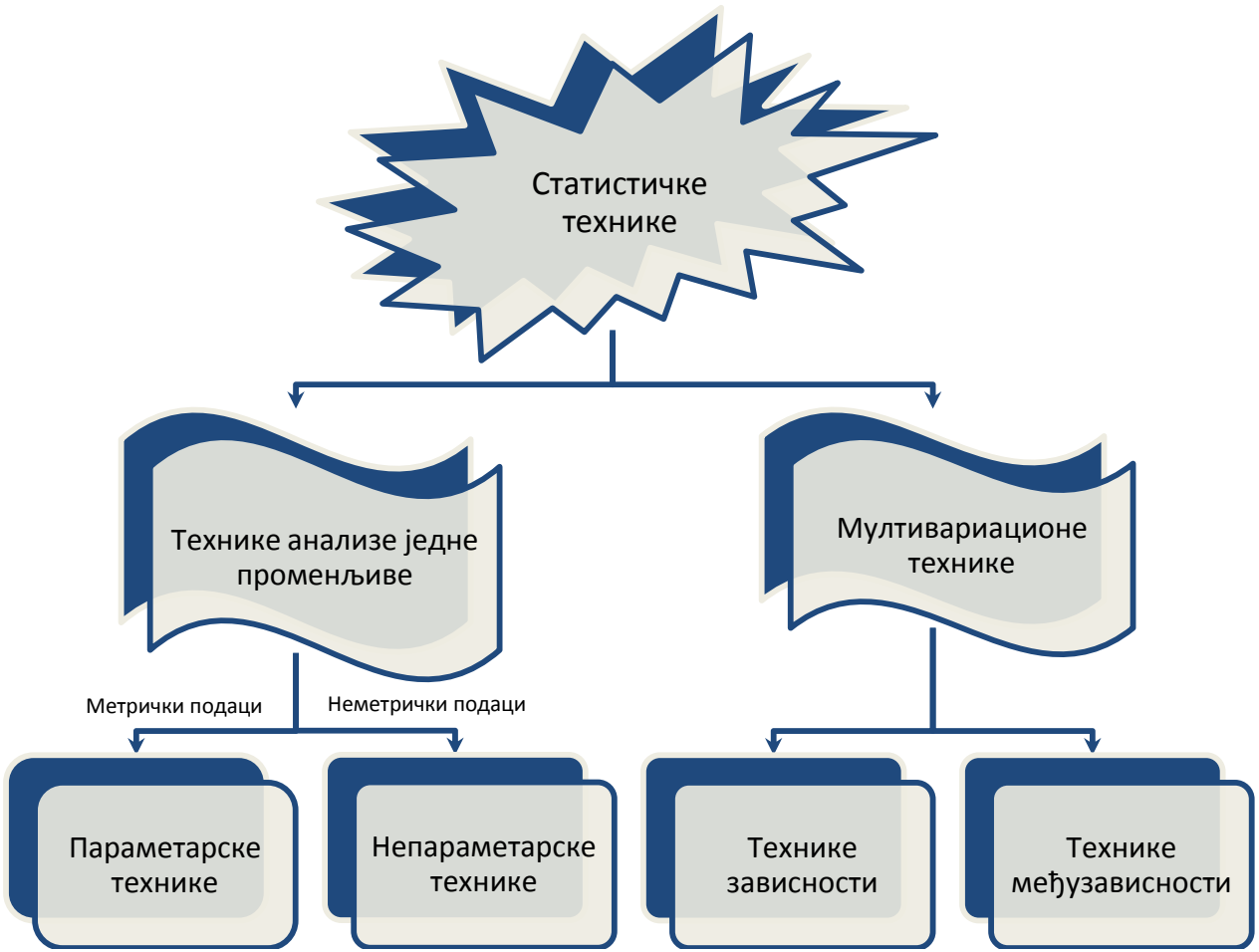
Рационална скала

- све карактеристике као интервална скала уз додатак апсолутне нулте тачке
- поређење две вредности преко количника (поређење различитих мерних јединица)
- примена највећег броја статистичких техника
- увек нумеричког типа

Примери:

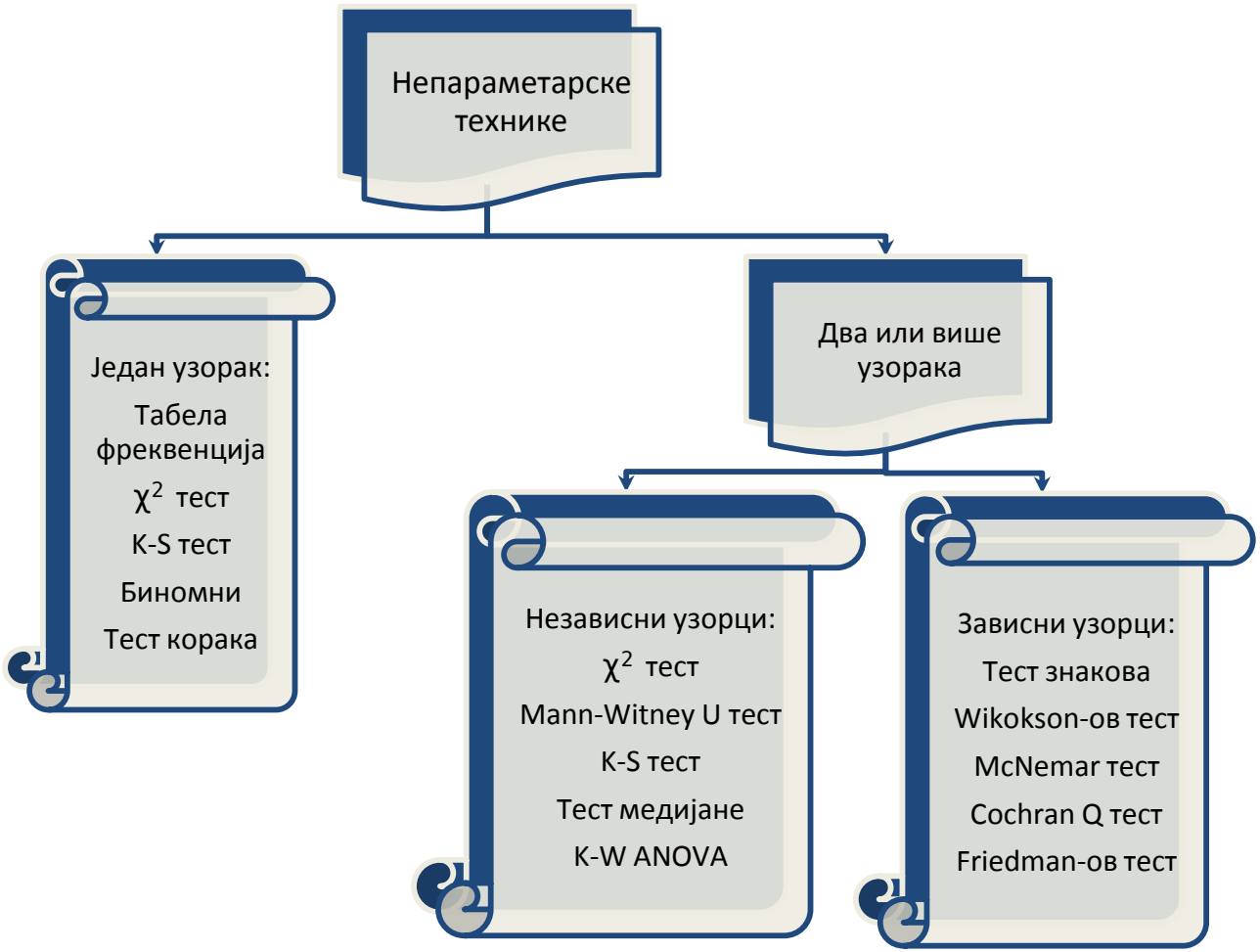
температура у Келвинима, висина, тежина, примања, време чекања на станици,...

Статистичке технике



Параметарске технике

Непараметарске технике



Мултивариационе технике

Мерне скале и статистичке технике

Номинална скала	
Фреквенције и мода	✓
Медијана и перцентили	✗
Сабирање и одузимање	✗
Средња вредности	✗
Стандардна девијација	✗
Однос и коефицијент варијације	✗
Хи-квадрат тест	✓
Биномни тест	✓

Ординална скала	
Фреквенције и мода	✓
Медијана и перцентили	✓
Сабирање и одузимање	✗
Средња вредности	✗
Стандардна девијација	✗
Однос и коефицијент варијације	✗
Корелација рангова	✓
Фридманова АНОВА	✓

Интервална скала	
Фреквенције и мода	✓
Медијана и перцентили	✓
Сабирање и одузимање	✓
Средња вредности	✓
Стандардна девијација	✓
Однос и коефицијент варијације	✗
Регресија и коефицијент корелације	✓
АНОВА	✓
Т-тест	✓
Факторска анализа	✓

Релациона скала	
Фреквенције и мода	✓
Медијана и перцентили	✓
Сабирање и одузимање	✓
Средња вредности	✓
Стандардна девијација	✓
Однос и коефицијент варијације	✓
Регресија и коефицијент корелације	✓
АНОВА	✓
Т-тест	✓
Факторска анализа	✓

Зашто је важно дефинисати мерне скале променљивих?

- Да би се примениле неке од статистика и метода у *SPSS*−у неопходно је да буду дефинисане мерне скале променљивих.
- Чак и када програм не захтева дефинисање мерне скале, то треба учинити да и се избегла евентуална неправилна обрада података.

Прелиминарна анализа података

- Процедуре за сумарно приказивање података
- Истраживачка анализа података

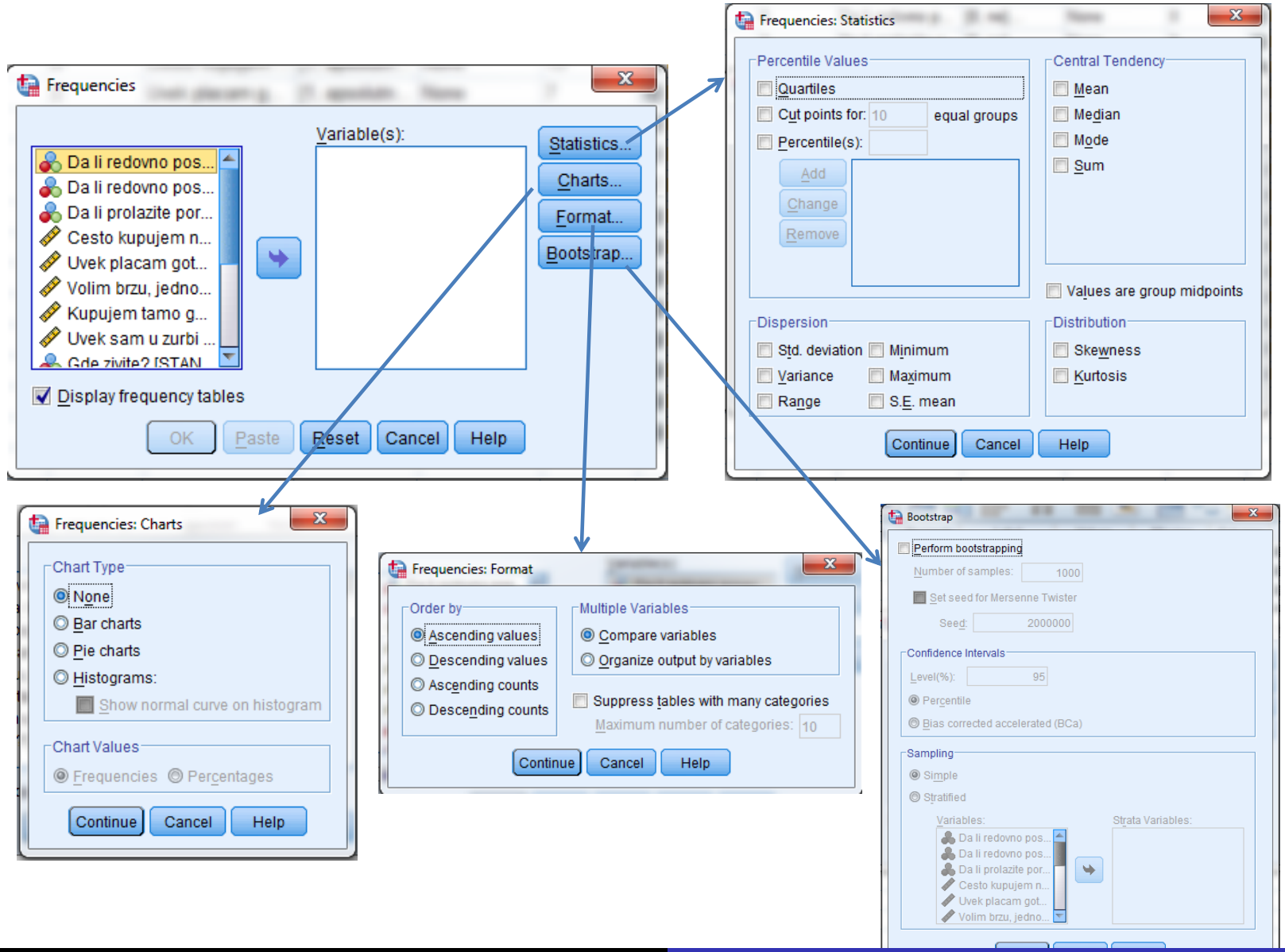
Analyze⇒Descriptive Statistics

Frequencies

- табела фреквенција за изабрану променљиву
- низ статистичких дескриптивних величина
- дијаграм

Табела фреквенција садржи:

- апсолутне фреквенције
- проценат
- валидни проценат
- кумулативни проценат


Frequencies
Variable(s):
Da li redovno pos...
Da li redovno pos...
Da li prolazite por...
Cesto kupujem n...
Uvek placam got...
Volim brzu, jedno...
Kupujem tamo g...
Uvek sam u zurbi ...
Gde zivite? [STAN
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help
Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis
Continue
Cancel
Help
Frequencies: Charts
Chart Type
None
Bar charts
Pie charts
Histograms:
Show normal curve on histogram
Chart Values
Frequencies
Percentages
Continue
Cancel
Help
Frequencies: Format
Order by
Ascending values
Descending values
Ascending counts
Descending counts
Multiple Variables
Compare variables
Organize output by variables
Suppress tables with many categories
Maximum number of categories: 10
Continue
Cancel
Help
Bootstrap
Perform bootstrapping
Number of samples: 1000
Set seed for Mersenne Twister
Seed: 2000000
Confidence Intervals
Level(%): 95
Percentile
Bias corrected accelerated (BCa)
Sampling
Simple
Stratified
Variables:
Strata Variables:

Пример

Frequency Table

Cesto kupujem na rasprodaji

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	apsolutno se ne slazem se	16	9,9	9,9	9,9
	ne slazem se	35	21,6	21,6	31,5
	neutralan	56	34,6	34,6	66,0
	slazem se	34	21,0	21,0	87,0
	apsolutno se slazem	21	13,0	13,0	100,0
	Total	162	100,0	100,0	

Gde zivite?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	svoja kuca	96	59,3	59,3	59,3
	rentira	66	40,7	40,7	100,0
	Total	162	100,0	100,0	

Kakav je Vas radni status?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	radi puno radno vreme	77	47,5	47,5	47,5
	radi pola radnog vremena	40	24,7	24,7	72,2
	penzioner/ne radi	45	27,8	27,8	100,0
	Total	162	100,0	100,0	

Пример са недостајућим вредностима

Da li prolazite pored marketa1 i marketa2 kada idete/vracate se sa posla?

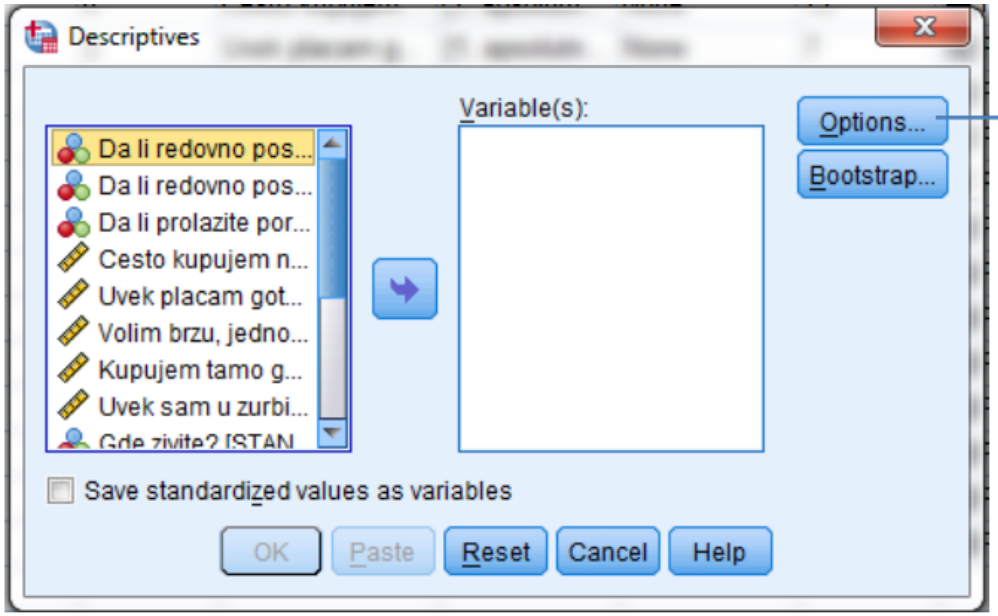
N	Valid	133
	Missing	29

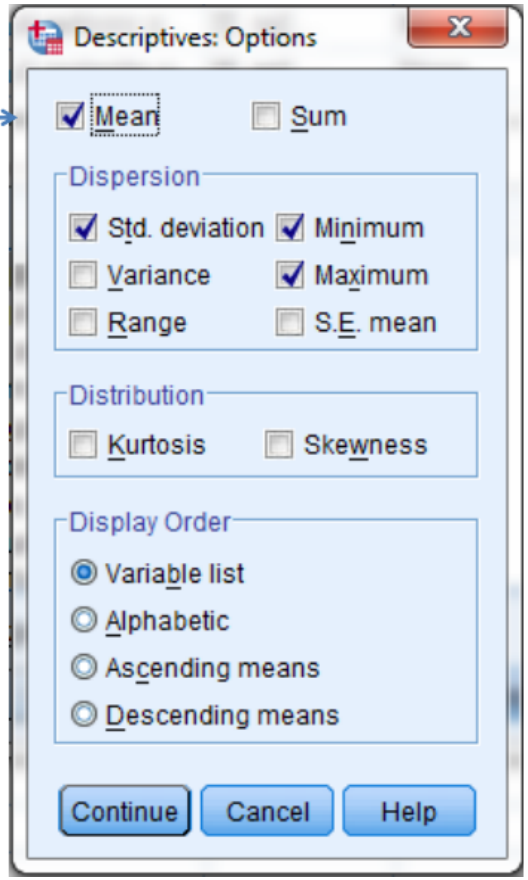
Da li prolazite pored marketa1 i marketa2 kada idete/vracate se sa posla?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ne	25	15,4	18,8	18,8
	da	108	66,7	81,2	100,0
	Total	133	82,1	100,0	
Missing	System	29	17,9		
Total		162	100,0		

Descriptives

- *Variable* - променљива(е) непрекидног типа
- *Options* - избор статистике централне тенденције, дисперзије, облика расподеле и редоследа приказивања резултата
- *Save standardized values as variables* - добија се нова променљива са стандардизованим вредностима





Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Prosecna vrednost kupovine u marketu1 u poslednjih nedelju dana	162	,00	18000,00	3883,7654	2488,04621
Valid N (listwise)	162				

Пример 1

Изабрати подниз запослених који имају садашњу плату мању од 20000 и радни стаж већи од 5 година, а затим приказати колико има таквих запослених по категорији посла. Користити базу *baza.sav*.

Пример 1

Изабрати подниз запослених који имају садашњу плату мању од 20000 и радни стаж већи од 5 година, а затим приказати колико има таквих запослених по категорији посла. Користити базу *baza.sav*.

- *Data* ⇒ *Select Cases*
- *Analyze* ⇒ *Descriptive Statistics* ⇒ *Frequencies* за променљиву *kat_posla*

Statistics

Kategorija zaposlenih

N	Valid	190
	Missing	0

Kategorija zaposlenih

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	sluzbenik	129	67,9	67,9	67,9
	obuceni sluzbenik	28	14,7	14,7	82,6
	sluzbenik obezbedjenja	27	14,2	14,2	96,8
	sluzbenik sa fakult. diplomom	3	1,6	1,6	98,4
	posebna grupa sluzbenika	3	1,6	1,6	100,0
	Total	190	100,0	100,0	

Пример 2

Колико има запослених чија се садашња плата налази у интервалу $[21000, 28000]$? Користити базу *baza.sav*.

Пример 2

Колико има запослених чија се садашња плата налази у интервалу $[21000, 28000]$? Користити базу *baza.sav*.

- *Transforme* $\Rightarrow$ *Compute Variable* функција *RANGE*(*splata*, 21000, 28000)
- *Analyze* $\Rightarrow$ *Descriptive Statistics* $\Rightarrow$ *Frequencies*

Statistics

splata_1

N	Valid	474
	Missing	0

splata_1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	431	90,9	90,9	90,9
	1,00	43	9,1	9,1	100,0
	Total	474	100,0	100,0	

Пример 3

Приказати просечну старост 20% запослених са најнижом садашњом платом, као и просечну старост 20% запослених са највишом садашњом платом. Користити базу *baza.sav*.

Пример 3

Приказати просечну старост 20% запослених са најнижом садашњом платом, као и просечну старост 20% запослених са највишом садашњом платом. Користити базу *baza.sav*.

- *Transforme* $\Rightarrow$ *Visual Bining*
- *Data* $\Rightarrow$ *Split File*
- *Data* $\Rightarrow$ *Select Cases*
  If : $sadasnjaPlata <= 1 | sadasnjaPlata >= 5$
- *Analyze* $\Rightarrow$ *Descriptive Statistics* $\Rightarrow$ *Descriptive* за променљиву *starost*

Напомена

Тражене две групе радника се могу формирати кроз процедуру
Anayze ⇒ Descriptive Statistics ⇒ Frequencies ⇒ Statistics *у опцији* Percentiles *за променљиву* splata.

Descriptive Statistics

Sadasnja plata (Binned)		N	Minimum	Maximum	Mean	Std. Deviation
<= 9180	Starost radnika	96	23,00	64,50	44,0341	15,28483
	Valid N (listwise)	96				
16441+	Starost radnika	94	27,50	56,67	34,6990	6,27587
	Valid N (listwise)	94				

Пример 4

Применом одговарајућих процедура одговорити на следећа питања:

a) Да ли жене или мушкарци у просеку користе интернет више часова недељно?

b) Код ког типа интернет корисника су жене (односно мушкарци) највише заступљени?

c) Приказати просечне вредности става испитаника према електронском маркетингу, електронској трговини и електронском банкарству. Урадити посебно за жене и мушкарце.

d) Колико часова недељно у просеку група верних корисника користи интернет?

Користити базу 2.

a) *Data* ⇒ *Split File* за променљиву *pol*, затим *Anayze* ⇒ *Descriptive Statistics* ⇒ *Descriptives* за променљиву *casovi*.

Descriptive Statistics

pol		N	Minimum	Maximum	Mean	Std. Deviation
muski	X2-koriscenje Interneta_cas/nedelja	50	3	36	14,34	7,649
	Valid N (listwise)	50				
zenski	X2-koriscenje Interneta_cas/nedelja	50	2	36	10,76	7,747
	Valid N (listwise)	50				

b) *Analyze* ⇒ *Descriptive Statistics* ⇒ *Frequencies* за променљиву *Tip_Internet_korisnika*.

Tip_Internet_korisnika

pol			Frequency	Percent	Valid Percent	Cumulative Percent
muski	Valid	slab korisnik	9	18,0	18,0	18,0
		srednji korisnik	21	42,0	42,0	60,0
		veliki korisnik	20	40,0	40,0	100,0
		Total	50	100,0	100,0	
zenski	Valid	slab korisnik	22	44,0	44,0	44,0
		srednji korisnik	14	28,0	28,0	72,0
		veliki korisnik	14	28,0	28,0	100,0
		Total	50	100,0	100,0	

c) *Analyze* ⇒ *Descriptive Statistics* ⇒ *Descriptive* за променљиве *el_marketing*, *el_tregovina*, *el_bankarstvo*.

Descriptive Statistics

pol		N	Minimum	Maximum	Mean	Std. Deviation
muski	X7-stav prema elektronskom marketingu	50	1	7	4,28	1,874
	X8-stav prema elektronskoj trgovini	50	1	7	4,16	1,476
	X9-stav prema elektronskom bankarstvu	50	1	7	3,40	1,616
	Valid N (listwise)	50				
zenski	X7-stav prema elektronskom marketingu	50	1	7	4,34	1,547
	X8-stav prema elektronskoj trgovini	50	2	7	5,06	1,346
	X9-stav prema elektronskom bankarstvu	50	1	8	3,50	1,515
	Valid N (listwise)	50				

c) *Data* ⇒ *Select Cases* за услов *Tip_Internet_korisnika* = 3 и *Anayze* ⇒ *Descriptive Statistics* ⇒ *Descriptive* за променљиву *Casovi*.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
X2-koriscenje Interneta_cas/nedelja	34	16	36	21,44	5,383
Valid N (listwise)	34				

Пример 5

Применом одговарајућих процедура одговорити на следећа питања:

a) У којој категорији посла има највише (односно најмање) запослених.

b) Приказати просечно радно искуство запослених за сваку категорију посла посебно.

c) Колико процената запослених ради у прве две категорије посла (приказати посебно за мушкарце и жене)?

Користити базу (*baza.sav*).

a) Analyze⇒Descriptive Statistics⇒Frequencies за променљиву kat_posla

a) Analyze⇒Descriptive Statistics⇒Frequencies за променљиву kat_posla

b) Поделити датотеку према категоријама посла: Data⇒Split File, а затим применити процедуру Analyze⇒Desriptive Statistics⇒Descriptive за променљиву r_staz

a) Analyze⇒Descriptive Statistics⇒Frequencies за променљиву kat_posla

b) Поделити датотеку према категоријама посла: Data⇒Split File, а затим применити процедуру Analyze⇒Desriptive Statistics⇒Descriptive за променљиву r_staz

c) Поделити датотеку према полу: Data⇒Split File и приказати табелу фреквенција за променљиву kat_posla

Хвала на пажњи!