Глава 1
Основи дескриптивне статистике

1. Уводни појмови

Реч статистика има двојако значење. У свакодневном животу под статистиком често подразумевамо произвољне нумеричке податке. Читајући дневну штампу, слушајући радио или гледајући телевизију често се сусрећемо са разноврсним статистичким подацима. Обично то нису појединачни подаци, већ читаве серије података, које представљају нумеричке вредности појединих величина којима се описују неки аспекти природних или друштвених појава, које су предмет истраживања.

Друго значење речи статистика је назив за научну дисциплину. Статистика као научна дисциплина обухвата научне методе који се користе за прикупљање, приказивање, анализу, интерпретацију података и доношење статистичких закључака. Одлуке које се доносе на основу статистичких метода се зову *процене* и *прогнозе*.

Статистика као научна дисциплина има два аспекта — теоријски и примењени. *Теоријска* или *математичка статистика* се бави развојем, извођењем и доказивањем статистичких теорема, метода, формула, правила и закона.

Примењена статистика примењује знања теоријске статистике у решавању реалних задатака. Примењена статистика се дели на *дескриптивну* и *инференцијалну статистику* (*статистику закључивања* или *индуктивну статистику*).

Серије података које добијамо прикупљањем информација о неком објекту нашег истраживања су обично велике и непрегледне. *Дескриптивна статистика* се састоји од метода за прикупљање, сређивање, приказивање и описивање података помоћу табела, графикона и сумарних показатеља.

Основни објекат на коме се нека појава статистички посматра назива се *основни скуп* или *популација*. Популација се састоји од свих *елемената* (*јединица посматрања*) који су датим статистичким истраживањем, у принципу, обухваћени, а који у "већој или мањој мери" поседују својство о коме желимо да, оно што је за нас релевантно у датом истраживању, закључимо. Фраза 'у принципу' у датој дефиницији захтева мало појашњење у следећем смислу. Произвољни непразни подскуп основног скупа се зове *узорак*. Најчешће имамо случај да дато статистичко истраживање обухвата, фактички, само део основног скупа, тј. неки узорак из дате популације, на основу чијих резултата изводимо закључак о читавој популацији.

Дакле, закључак о основном скупу, у смислу посматраног својства, се готово увек доноси на основу дела тог скупа, односно на основу узорка.

Елементи основног скупа могу бити произвољне природе. Својства основног скупа која посматрамо зову се *променљиве*; променљиве обично означавамо великим словима латинице $X, Y, Z, \ldots$.

Инференцијална статистика обухвата статистичке методе које примењујемо да бисмо на основу информација добијених о елементима узорка дошли до закључака о својствима (карактеристикама) основног скупа.

Нека је Ω основни скуп, а X његово својство које истражујемо. Ако је $\omega \in \Omega$ неки елемент датог основног скупа, онда са нумеричком вредношћу $X(\omega)$ означимо "количину" посматраног својства коју поседује елемент ω. Дакле, за нас је променљива X функција $X\colon \Omega \to \mathbf{R}$.

Пример 1. На пример, ако нас интересује:

а) проценат укупног броја становника Београда који је старији од 20 година;

б) зараде свих запослених неког предузећа;

в) цене свих аутомобила које продаје фирма HitAuto,

онда је у првом случају основни скуп сви становници Београда, у случају б) сви запослени посматраног предузећа, а у последњем случају сви типови аутомобила које продаје фирма HitAuto. □

Пример 2. На пример, треба наћи просечну дужину стопала свих одраслих мушкараца који живе на Куби. Тако нешто би, рецимо, интересовало руководство фирме тек изграђене фабрике обуће у граду Санта Клара. Приметимо да дужина стопала, као и многе друге величине у природи, зависе од много случајних фактора и слободно можемо претпоставити да она има тзв. нормалну или Гаусову расподелу, а за ову расподелу важи тзв. емпиријско правило — 99.7% вредности се налази у 3σ-околини средње вредности, где је σ стандардна девијација одговарајуће нормалне расподеле.

Било би готово немогуће (временски и финансијски нерационално) извршити мерење стопала свих одраслих мушкараца Кубе. Стога руководство фирме одлучује да поступи другачије. Фирма наручује статистичко истраживање које ће на основу дела основног скупа, тј. узорка, утврдити како се тражена величина понаша на читавом основном скупу. Ако је узети узорак репрезентативан за посматрану величину, онда је добијена информација о средњој вредности посматране величине за елементе узорка приближно једнака траженој средњој вредности. □

Да би се на основу узорка могло закључити што тачније о карактеристикама основног скупа, узорак мора бити *репрезентативан*: узорак мора у највећој мери да одражава карактеристике основног скупа.

Пример 3. Ако желимо на основу узорка да сазнамо просечан приход породица Београда, онда узорак мора да садржи породице из разних категорија готово у истој сразмери која постоји имеђу њих у основном скупу. □

Узорак може бити *случајан* и *неслучајан*. Ако приликом избора узорка сваки елемент основног скупа има унапред познату (позитивну) вероватноћу да буде изабран у узорак, онда је то *случајан узорак*. Случајни узорак је обично и репрезентативан. Ако неки елементи основног скупа немају шансу да буду изабрани у узорак (вероватноћа њиховог избора у узорак је једнака 0), онда је узорак *неслучајан*; уместо неслучајан узорак користи се и термин *намеран* узорак.

Пример 4. На пример, претпоставимо да нас интересује просечна оцена с којом су завршили прву годину студирања студенти Бета универзитета који су на време завршили прву годину студирања. До тог податка желимо да дођемо тако што ћемо узети узорак од 30 студената који су на време завршили прву годину, на пример, у последњих 5 година. До узорка можемо доћи тако што ћемо исписати на цедуљице имена свих студената који су завршили прву годину на време у том периоду, убацити ове цедуљице у "шешир" и на случајан начин извући из њега 30 имена. А можемо и тако, што ћемо формирати списак свих студената, о којима је реч, а затим узети са тог списка првих 30 имена. У првом случају се добија случајан узорак, а у другом случају је узорак неслучајан, јер неки од студената немају уопште шансе да буду укључени у узорак. □

Убудуће, број елемената основног скупа означаваћемо са N, а број елемената у узетом узорку, тзв. *дужину узорка*, са n.

Углавном посматрамо тзв. узорак без понављања, мада се у теорији посматра и узорак са понављањем (о овоме ће касније бити више речи). Ако посматрамо узорак без понављања, онда је он, просто, сваки непразни подскуп основног скупа. Дакле, ако основни скуп има N елемената, онда постоји $2^N - 1$ различитих узорака. Број узорака дужине n, где је $n \leq N$, је, тада, једнак

$$C_N^n = \binom{N}{n} = \frac{N(N-1)\dots(N-n+1)}{n!}.$$

Ако сви узорци исте дужине које бирамо из основног скупа имају исту вероватноћу да буду изабрани, онда се сам поступак бирања зове *избор простог случајног узорка*, а узорак *простим случајним узорком.*

Као примере неслучајних узорака поменимо *погодан узорак* и *узорак заснован на субјективном суду истраживача*. Код погодног узорка бирају се најдоступнији елементи основног скупа да би се брзо дошло до резултата (на пример, ако се неко истраживање тиче потрошача, ми можемо да се ограничимо само једним тржним центром). Овакав узорак обично није репрезентативан. Код узорка заснованог на субјективном суду истраживача елементи из основног скупа бирају се на основу процене и претходног знања истрживача. Такав узорак може бити репрезентативан, али вероватноћа да је он такав је веома мала.

За доношење правилних прогноза и процена важно је да располажемо поузданим подацима. Извори поузданих података се могу поделити у две основне категорије: *примарни* и *секундарни* извори података.

Примарни подаци су подаци које сам истраживач (организација) прикупља по први пут за потребе неког конкретног истраживачког пројекта.

Три основна начина за прикупљање примарних података су: *посматрање*, *испитивање* и *експериментисање*.

Секундарни подаци су информације које већ негде постоје и прикупљени су за неку другу сврху. Секундарни извори података могу бити *интерним* или *екстерним*.

Интерни секундарни подаци су, на пример, подаци фирме које она поседује о својим запосленима или књиговодствени подаци, извештаји о ранијим истраживањима, извештаји о продаји, известаји о трошковима маркетинга, извештаји о поруџбинама, извештаји о рекламацијама и сл.

Извори екстерних секундарних података су: публикације националних и међународних организација, публикације статистичких завода, извештаји слузбе за платни промет, извештаји привредних комора, новине и стручни часописи приручници, каталози и адресари, базе података и сл.

Прикупљање података о елементима основног скупа или узорка може се спровести путем анкете. У анкети се подаци прикупљају о елементима основног скупа или узорка, без контроле фактора који могу да утичу на карактеристике које нас интересују или резултате анкете. Ако је анкетом обухваћен сваки елемент основног скупа, онда се таква анкета зове *попис*. Ако се анкета врши само над елементима узорка, онда се она зове *узорачка анкета*.

Три су главна разлога зашто узорачкој анкети дајемо предност у односу на попис:

1) <u>Време</u>. Основни скуп је понекад тако велик да резултати пописа могу чак, током пописа, да постану неактуелни.

2) <u>Трошкови</u>. Што је основни скуп већи то су трошкови везани за извођење пописа већи.

3) <u>Немогућност спровођења пописа</u>. Није могуће идентификовати све елементе основног скупа и доћи у контакт са њима (види пример са бескучницима).

У принципу постоје три технике како се анкета може обавити: *интервју*, *телефонским путем* и *путем поште*.

Интервју има предност због великог одзива и квалитета добијених одговора, али је најскупља техника и захтева највише времена од осталих техника.

Телефонска анкета има велики одзив, али неки људи не воле да их узнемиравају телефоном, а и они који немају телефон изостављени су из ње.

Анкета која се спроводи *путем поште* је најефтинији метод, али је одзив обично низак — многи људи не враћају упитнике.

Резултати добијени применом узорачке анкете могу да садрже два типа грешака: *случајне* (или *узорачке* грешке) и *неслучајне* (*неузорачке* или *систематске* грешке).

Случајна грешка је разлика између резултата добијеног узорачком анкетом и одговарајућег резултата који би смо добили да смо применили попис.

Грешке које се јављају при прикупљању, бележењу и формирању табела података зову се *неслучајне грешке*. Неслучајне грешке могу бити: *грешка због погрешно изабраног оквира узорка*, *грешка због одсуства одговора*, *грешка у одговору* и *грешка добровољног одговора*.

Листа елемената основног скупа која се користи при избору узорка зове се *оквир узорка* (на пример, ако се при избору узорка користимо телефонским именником, онда је листа имена у њему оквир узорка). Ако оквир узорка није репрезентативан за основни скуп, тада долази до *грешке због погрешно изабраног оквира узорка* (у примеру са телефонским именником, због лоше изабраног оквира из узорка су унапред искључени, на пример, сви бескучници, али исто тако и сви они који из разних разлога не желе да им се број телефона нађе у телефонском именику).

Грешка која се јавља у случају када многи испитаници укључени у анкету нису дали одговор зове се *грешка због одсуства одговора* (такве грешке обавезно прате анкете путем поште, где је примећено да обично људи са ниским и високим примањима просто не враћају упитнике).

Грешка у одговору се јавља када испитаници обухвачени анкетом не дају тачан одговор (узрок за тако нешто може бити нејасна формулација питања, питања могу бити шкакљива за испитаника, а могу зависити и од пола и расе испитивача).

Грешка добровољног одговора се јавља када се анкета не спроводи на случајном узорку, већ се упитник објави у неком часопису или преко неке телевизијске станице (такве анкете се зову и *псеудоанкетама*), па се читаоци позову да попуне и пошаљу поштом или се јаве телефоном. Обично се јављају само они који имају јако изражен став према питањима која се дотичу анкетом, а и одговор може бити повезан са новчаним издацима (рецимо, треба платити цену телефонског разговора). Овакав узорак обично није репрезентативан, јер учешће је добровољно.

Постоји више техника за добијање случајног узорка. Већ смо видели како се прости случајни узорак може добити методом случајног извлачења без понављања из "шешира" (овај шешир може бити списак из кога се уз помоћ генератора случајних бројева извлаче имена).

Други тип случајног узорка је такозвани *систематски случајан узорак*. Објаснимо га на једном примеру. На пример, из популације од 100000 људи треба случајно изабрати 100. Формирајмо списак свих елемената популације и поделимо га на сто блокова 100000/100 = 1000 особа (први блок садржи све особе са списка од прве до особе са редним бројем 1000, други блок од особе са редним бројем 1001 до особе са редним бројем 2000, итд.). Случајно из првог блока изберемо особу, на пример, особу под редним бројем n. Тада се наш узорак састоји од особа из списка са редним бројевима

$$n, 1000 + n, 2000 + n, \ldots, 99000 + n.$$

Често се дешава да је природа проблема при статистичком истраживању таква да зависи од два или више изражених фактора. Ако су ти фактори добро дефинисани, погодно је применити технику која се базира на тзв. *квота узорку*. У складу са уоченим факторима прво извршимо разбијање основног скупа на дисјунктне класе. Свака од ових класа се зове *стратум*. Затим утврдимо процентуално учешће сваког од ових стратума у основном скупу, а затим из сваког од стратума изаберемо узорак и објединимо све ове узорке у један. При томе захтевамо да је процентуално учешће сваког *стратум-узорка* у обједињеном узорку једнако процентуалном учешћу одговарајућег стратума у основном скупу. Ако је последње испуњено онда се добијени узорак зове *квота узорак*, а ако је сваки од појединих узорака који учествују у финалном узорку био случајан, онда се добијени узорак зове *стратификован случајан узорак*. Фактори који утичу на потребу за разбијањем основног скупа могу бити: пол, приходи, трошкови, образовање, раса, запосленост, итд.

Пример 5. На пример, често су политички аналитичари наглашавали, када се у другом кругу председничких избора водила битка између два председничка кандидата — X и Y, да у случају када би гласао само женски део популације председник би био Y, а у случају када би гласао само мушки део победио би X. Како је жена, у основној популацији, приближно 52%, а мушкараца 48%, то ако желимо да испитамо јавно мњење на стратификованом случајном узорку од 1000 људи, неопходно је случајно изабрати 520 жена и 480 мушкараца. □

Још један тип случајног узорка је тзв. *кластер узорак*, који се примењује, нарочито, у случају када је основни скуп распрострањен на великој географској површини. Овакав узорак добијамо тако што основни скуп поделимо на различите географске групе или *кластере* (кластер мора бити репрезентативан за дату популацију). Затим изаберемо случајан узорак кластера, а онда из сваког изабраног кластера по један случајан узорак. Сада, објединивши тако добијене узорке, добијамо *кластер узорак*.

2. Серије података. Типови променљивих

Опсервација или *податак* је вредност променљиве која се односи на један елемент, тј. вредност коју посматрана променљива узима у једној тачки основног скупа. *Серија података* је скуп података који се односи на једну или више променљивих.

Пример 6. На пример, погледајмо доњу таблицу (подаци из Wikipedia-е о доказаним резервама нафте на дан 12. март 2013.). Променљива која је "предмет" ове таблице је 'резерве нафте' (резерве су изражене у милиардама барела). Јединица посматања је свака конкретно наведена земља у таблици, опсервација или податак је количина нафте коју посматрани елемент (земља) има, а читава таблица представља серију података.

Земља	**Резерве нафте**
Саудијска Арабија	266.08
Канада	178.60
Ирак	143.50
Иран	137.60
Кувајт	104.00
УАЕ	97.80
Венецуела	87.04
Русија	79.00
Либија	41.46
Нигерија	36.22

Подаци прикупљени о различитим елементима основног скупа или различитим елементима узорка у истом временском тренутку, или у истом временском периоду се зову *структурна серија* (*подаци пресека* или *упоредни подаци*).

Серија података прикупљених за исту јединицу посматрања о једној истој променљивој у различитим временским тренуцима или временским периодима се зове *временска серија*.

Негруписани подаци (*сирови подаци*) су подаци записани редоследом којим се прикупљају, пре него што се уреде по величини, или групишу.

Променљива може бити *квантитативна* или *квалитативна*. Вредности квантитативних променљивих, квантитативни подаци, се изражавају реалним бројевима, а вредности квалитативних променљивих, квалитативни подаци, нису, у суштини, нумерички подаци (они могу бити кодирани бројевима, али се с њима не може оперисати као са бројевима), него различити модалитети неког атрибута (на пример, то може бити марка аутомобила, боја косе, пол, итд.). *Категорију* чине сви елементи основног скупа који поседују исти модалитет. За квалитативне променљиве се још користи термин *категоријска*, односно, *атрибутивна* променљива.

Квантитативна променљива, са своје стране може бити *прекидна* (или *дискретна*) (на пример, број аутомобила, саобраћајних незгода, итд.) и *непрекидна* (на пример, дужина, старост, висина, тежина, време, итд.). Дискретна променљива узима највише пребројиво вредности, а непрекидна непребројиво (то су обично све реалне вредности које припадају неком интервалу реалних бројева).

3. Сређивање и графичко приказивање квалитативних података

Фреквенција (*учесталост*) одређене категорије је број елемената основног скупа који припадају тој категорији. *Расподела фреквенција* показује како су фреквенције распоређене по различитим категоријама и даје се, обично, *табелом расподеле фреквенција* или *табелом фреквенција*. Објаснимо ове појмове једним примером.

Пример 7. Једним истраживањем, којим се хтело утврдити за које врсте занимања ће се определити студенти после дипломирања, обухваћено је 100 студената једног факултета. Резултати одговарајуће анкете представљени су следећом таблицом:

Врста запослења	**Број студената (f)**
Државне фирме и установе	16
Приватне фирме/компаније	44
Сопствене фирме	17
Државни/локални органи власти	23

Основни скуп у нашем примеру је скуп 100 анкетираних студената. Случајна променљива је врста запослења. У првој колони таблице су дате могуће ‘вредности’ које ова променљива може узети, а које у случају квалитативних променљивих зовемо категоријама или модалитетима. У другој колони (колони фреквенција) се налазе одговарајуће фреквенције ових категорија. На пример, из табеле можемо прочитати да се 16 студената определило да после дипломирања ради у државним фирмама и установама. Дата таблица и јесте табела расподеле фреквенција за дату променљиву. Приметимо да се фреквенција означава са f, односно, фреквенција i-те категорије се означава са f_i.

Релативна фреквенција дате категорије добија се када се њена фреквенција подели са сумом свих фреквенција, тј важи следећа формула:

$$\textbf{релативна фреквенција } i\textbf{-те категорије} = \frac{f_i}{\sum f}\,.$$

Учешће (*проценат*) i-те категорије добија се по следећој формули:

$$\textbf{учешће } i\textbf{-те категорије} = \\ = \textbf{релативна фреквенција } i\textbf{-те категорије} \times 100\,.$$

Учешћем се се одређује удео, изражен у процентима, дате категорије у основном скупу. *Расподела релативних фреквенција* [*процентуална расподела*] показује како су релативне фреквенције [учешћа] распоређене [распоредјена] по различитим категоријама. Ове расподеле се обично задају одговарајућим табелама. Илуструјмо уведене појмове следећим примером.

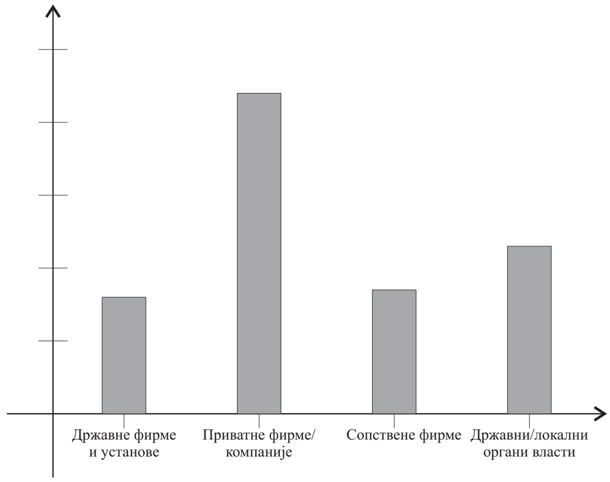
Пример 8. Ако погледамо анкету из последњег примера, онда ће одговарајућа табела релативних фреквенција, односно, табела процентуалне расподеле, имати следећу форму (обе ове табеле дајемо заједно, у облику једне табеле):

Врста запослења	**Релативна фреквенција**	**Учешће**
Државне фирме и установе	0,16	16%
Приватне фирме/компаније	0,44	44%
Сопствене фирме	0,17	17%
Државни/локални органи власти	0,23	23%

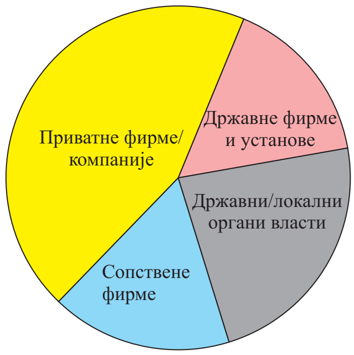
Квалитативни подаци се могу графички приказати уз помоћ два типа дијаграма — *штапићастог дијаграма* и *структурног круга*.

Штапићасти дијаграм је графикон који се састоји од стубића једнаке ширине, чије основице леже на x-оси и центри тих основица чине систем еквидистантних тачака (растојање између сваке две суседне тачке овог система је једнако некој унапред фиксираној вредности), који својим висинама представљају фреквенције различитих категорија. Свакој од категорија припада свој стубић и између стубића се оставља обавезно слободни простор. Овим дијаграмом могу се представити не само фреквенције, него и релативне фреквенције и учешћа различитих категорија посматране квалитативне промењиве.

Доле дати штапићасти дијаграм представља фреквенције различитих категорија случајне промењиве из примера 7.



Структурни круг (*пита*) је круг подељен на делове (секторе), сваки од делова представља релативну фреквенцију или учешће одговарајуће категорије у основном скупу. Величине тих делова су у корелацији са одговарајућим величинама релативних фреквенција [учешћа]. На слици здесна дат је структурни круг учешћа различитих категорија за случајну променљиву из примера 7.



4. Сређивање и графичко приказивање квантитативних података

Веома често се при неком статистичком истраживању барата са великом количином сакупљених нумеричких података. Ради лакше обраде ових података згодно их је груписати на класе "веома сличних" података. "Критеријум сличности" података се састоји у следећем. Уочавамо неки низ дисјунктних бројних интервала (између суседних интервала може постојати "празан простор", али исто тако уочени интервали могу налегати један на други) који својом унијом покривају (садрже) све уочене податке; ове бројне интервале зовемо *групним интервалима*. Дужине ових интервала могу бити различите, а ове дужине као и "локације" ових интервала могу зависити од много фактора, неки од њих могу бити и субјективне природе. У идеалном случају пожељно би било да дужина групних интервала буду што мање, а њихов број што већи. Али у том случају би се сва предност оваквог прилаза обради података изгубила, те је стога у пракси број групних интервала обично између 5 и 20.

Број групних интервала c обично се одређује на основу *Старџесовог правила*:

$$c = 1 + 3{,}3 \log N,$$

где се са N, као што смо рекли, означава укупан број података (број елемената основног скупа).

Уочени групни интервали, у принципу, могу бити различити по ширини, али се обично *ширина групних интервала* узима једнаком за све њих. И не само то, ако постоји празан простор (такође, интервал) између њих, он се такође узима једнаким за све парове суседних интервала. До ширине групних интервала долазимо тако што прво одређујемо тзв. *апроксимативну ширину групних интервала* из следеће формуле:

$$\textbf{апроксимативна ширина групних интервала} = \\ = \frac{\textbf{највећа вредност} - \textbf{најмања вредност}}{\textbf{број групних интервала}}$$

У формули *најмања* [*највећа*] *вредност* је вредност најмањег [највећег] податка. Овако добијену величину "заокружујемо" на неки цео број (овај цео број не мора бити цео број који је најближи добијеној вредности из горње формуле). Овакво заокруживање може незнатно да промени број групних интевала.

Како је сваки од групних интервала, у суштини, интервал реалних бројева, он има свој леви и десни крај који се редом зову *доња граница* и *горња граница* датог групног интервала.

Граничне вредности групних интервала су средине између горње границе неког групног интервала и доње границе наредног интервала. За уочени групни интервал прва слева [здесна] овако добијена вредност зове се његова *доња* [*горња*] *гранична вредност*, а обе ове вредности су *граничне вредности* посматраног групног интервала. Очигледно, да за први групни интервал постоји само горња гранична вредност, а за последњи групни интервал доња гранична вредност. Тада се *ширина* (одређеног групног интервала) израчунава по следећој формули:

$$\textbf{ширина групног интервала} = \\ = \textbf{горња гранична вредност} - \textbf{доња гранична вредност}.$$

Такође, уводи се и појам *средине групног интервала* следећом формулом:

$$\textbf{средина групног интервала} = \\ = \frac{\textbf{доња граница} + \textbf{горња граница}}{2}.$$

Фреквенција датог групног интервала је број података (одговарајућих елемената посматране популације или узорка) који припадају уоченом интервалу. Ако је расподела фреквенција нумеричких података представљена низом групних интервала и одговарајућим низом фреквенција за те интервале, онда се подаци приказани оваквом расподелом зову *груписани подаци*.

Објаснимо уведене појмове на следећа два примера.

Пример 9. Садржај једног статистичког истраживања биле су недељне зараде 100 запослених једне приватне фирме. Подаци су изложени у облику следеће табеле.

Недељна зарада	**Број запослених (f)**
401–600	9
601–800	22
801–1000	39
1001–1200	15
1201–1400	9
1401–1600	6

Групни интервали, а њих укупно, као што видимо, има 6, задати су првом колоном ове табеле. На пример, четврти групни интервал је дефинисан интервалом реалних бројева $[1001; 1200]$. Доња [горња] граница четвртог групног интервала је 1001 [1200], а доња [горња] гранична вредност четвртог групног интервала је 1000,5 [1200,5]. Ширина четвртог групног интервала (а, такође, и свих осталих) је $1200{,}5 - 1000{,}5 = 200$. Фреквенције (број запослених по групним интервалима) променљиве 'Недељна зарада' су задате другом колоном ове табеле, и, на пример, за четврти групни интервал она износи 15, тј. 15 запослених ове фирме има недељне зараде у интервалу од 1001 \$ до 1200 \$. □

Приметимо да је променљива '**недељна зарада'** променљива дискретног типа. Наиме, зараде су изражене у целим доларима. Обично, у случају дискретних променљивих се између групних интервала оставља празан простор, али ако су у питању променљиве непрекидног типа, као што је то случај у следећем примеру, онда групни интервали належу један на други.

Пример 10. На пример, погледајмо поново пример 6, и уочимо за дате податке 5 групних интервала. Тада је

$$\text{апроксимативна ширина групних интервала} = \frac{266{,}08 - 36{,}22}{5} = 45,97\,.$$

Приметимо, да када би смо заокружили добијену вредност на први већи цео број, а то је 46, онда се сви подаци не би нашли у пет групних интервала, стога заокружимо добијену вредност на 47. Дакле, узећемо 5 групних интервала ширине 47, а одговарајућа табела фреквенција изгледаће овако:

Резерве нафте	**Број земаља**
од 36 до < 83	3
од 83 до < 130	3
од 130 до < 177	2
од 177 до < 224	1
од 224 до < 271	1

Овде, слично као и у случају квалитативних променљивих, можемо дефинисати релативну фреквенцију и учешће, али сада појединих групних интервала (а, такође, и одговарајуће расподеле релативних фреквенција и учешћа) на следећи начин:

$$\textbf{релативна фреквенција } i\textbf{-тог групног интервала} = \frac{f_i}{\sum f}$$

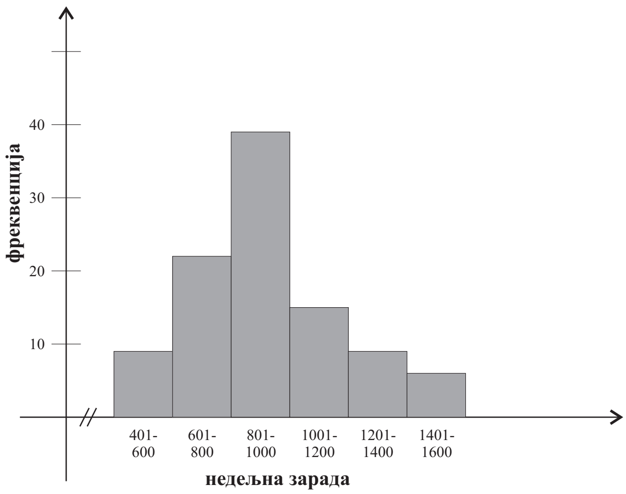
$$\textbf{учешће } i\textbf{-тог групног интервала} = \\ = \textbf{релативна фреквенција } i\textbf{-тог групног интервала} \times 100\,.$$

Пример 11. Погледајмо горе дати пример са недељним зарадама запослених. Одговарајуће табеле расподеле релативних фреквенција и процентуалне расподеле имају следећи облик (као и раније, обе табеле дајемо у оквиру једне):

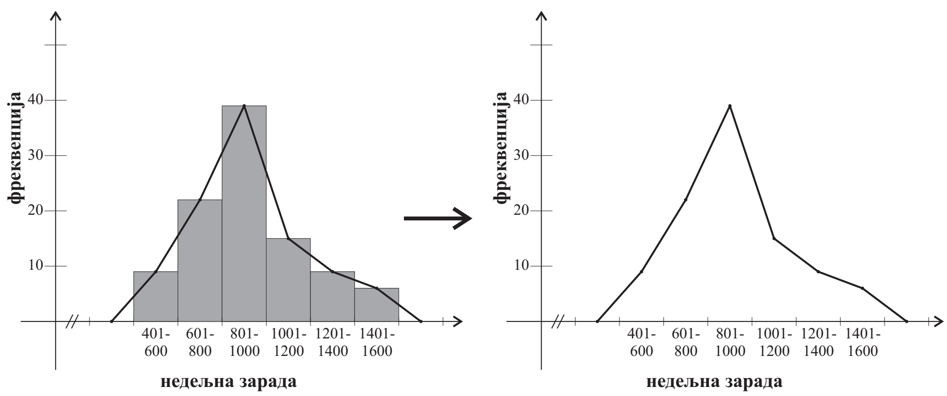
Недељна зарада	**Релативна фреквенција**	**Учешће**
401–600	$0,09$	9%
601–800	$0,22$	22%
801–1000	$0,39$	39%
1001–1200	$0,15$	15%
1201–1400	$0,09$	9%
1401–1600	$0,06$	6%

Груписани подаци се приказују обично уз помоћ *хистограма* и *полигона*.

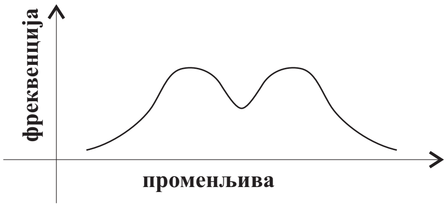
Хистограм фреквенција добијамо када над сваким групним интервалом, као основицом, конструишемо правоугаоник висине која је сразмерна фреквенцији посматраног групног интервала. За разлику од штапичастог дијаграма, код хистограма уочени правоугаоници належу један на други. Хистограм, очигледно, се може са истим успехом формирати како за релативне фреквенције, тако и за учешћа. На тај начин добијамо *хистограм релативних фреквенција* и *хистограм учешћа*. Одговарајућа скала фреквенција, релативних фреквенција или учешћа се у случају хистограма налази на y-оси. Хистограм фреквенција за груписане податке из примера 9 је



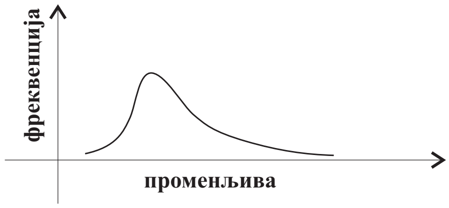
У зависности од тога шта је предмет нашег разматрања *полигон* може бити *полигон фреквенција*, *полигон релативних фреквенција*, или *полигон учешћа*. Да би смо добили *полигон фреквенција* поступамо на следећи начин. Полазимо од хистограма фреквенција и додајемо још по један групни интервал слеве и здесне стране од уочених групних интервала и формално над њима конструишимо правоугаонике висине 0. Затим повежимо све средине горњих основица правоугаоника тако допуњеног хистограма праволинијским одсечцима. Тако добијена изломљена линија и јесте полигон. Крајеви полигона, као што видимо, леже на x-оси. Слично, полазећи од хистограма релативних фреквенција [учешћа] добијамо полигон релативних фреквенција [учешћа]. Полигон фреквенција за груписане податке из примера 9 је



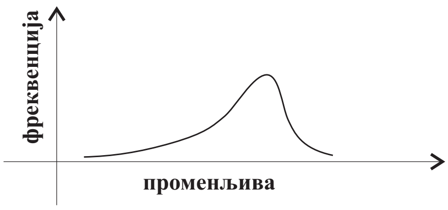
Претпоставимо да “имамо посла” са непрекидном променљивом и са веома дугачком серијом података. Ако повећавамо број групних интервала, и, у исто време, смањујемо њихове ширине (притом у мислима дозвољавамо да број групних интервала тежи ка бесконачности, а њихова ширина ка 0), ми се све више приближавамо, одговарајућим полигонима, ка једној глаткој континуираној (непрекидној) кривој, која се зове *крива расподеле фреквенција*, или *крива фреквенција*, или, просто, *теоријска крива*. Полигони, које добијамо у процесу те ‘мисаоне процедуре’ су, уствари, апроксимације овако добијене теоријске криве. У зависности од облика теоријске криве, односно, хистограма (ако је број групних интервала довољно велик), ове објекте класификујемо на симетричне, асиметричне удесно, асиметричне улево и униформне (правоугаоне).



Симетрична крива фреквенције



Крива фреквенције асиметрична удесно



Крива фреквенције асиметрична улево

Осим фреквенција групних интервала, често, се уочавају и тзв. њихове кумулативне фреквенције. *Кумулативна фреквенција* групног интервала приказује укупан број јединица посматрања које имају вредност испод горње границе сваког групног интервала. За сваки посматрани групни интервал, такође, се могу уочити:

$$\textbf{кумулативна релативна фреквенција } i\textbf{-тог интервала} = \\ = \frac{\textbf{кумулативна фреквенција } i\textbf{-тог интервала}}{\textbf{укупан број података}}$$

и

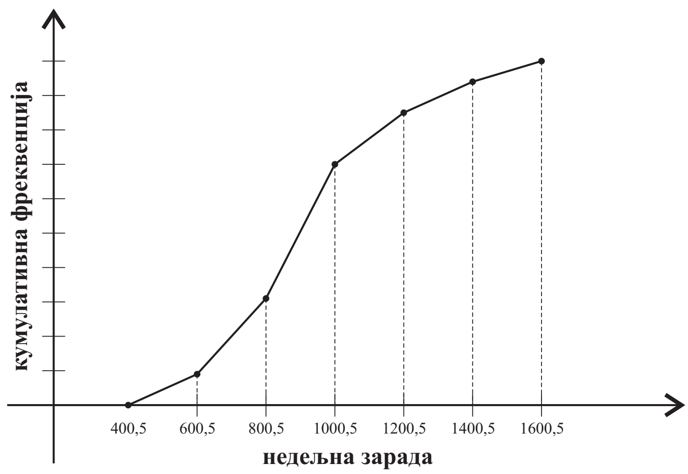
$$\textbf{кумулативно учешће } i\textbf{-тог интервала} = \\ = \textbf{кумулативна релативна фреквенција } i\textbf{-тог интервала} \times 100\,.$$

Као и раније, може се уочити расподела кумулативних фреквенција, односно, табела ове расподеле. Аналогно овоме, може се уочити расподела, односно, табела расподеле за кумулативне релативне фреквенције, односно, за кумулативна учешћа.

Пример 12. За пример са недељним зарадама одговарајућа табела расподеле кумулативних фреквенција има следећи облик:

Недељна зарада	**Број запослених** (f)
401–600	9
601–800	31
801–1000	70
1001–1200	85
1201–1400	94
1401–1600	100

Графички се расподела кумулативних фреквенција може представити тзв. кумулантом. *Кумуланта* или *огива* је крива која има облик полигоналне линије и која се добија на следећи начин. Нека су $a_1, a_2, \ldots, a_n$ горње граничне вредности уочених групних интервала, $f_1, f_2, \ldots, f_n$ одговарајуће њихове кумулативне фреквенције, а a_0 доња гранична вредност првог групног интервала. Повежимо тачке (a_i, f_i) и (a_{i+1}, f_{i+1}) праволинијским одсечцима за свако i, $i = 0, 1, 2, \ldots, n-1$; овде узимамо да је $f_0 = 0$. Добијена полигонална линија и јесте огива.



Постоји још неколико удобних начина за представљање квантитативних података међу којима се најчешће користи приказ у облику стабљике и листа или краће *стабло-лист дијаграм* и *тачкасти дијаграм.* Објаснимо ове начине представљања података следећим примерима.

Пример 13. Један од колоквијума из Пословне статистике који је одржан на Факултету за трговину и банкарство Бета универзитета полагало је 15 студената. Њихов успех, изражен освојеним поенима, је дат у облику следећег низа:

21, 11, 15, 13, 24, 21, 23, 5, 17, 18, 21, 9, 12, 22, 19

Дати низ података можемо представити у облику следећег дијаграма

2	1	4	1	3	1	2	
1	1	5	3	7	8	2	9
0	5						

У претходном примеру првом колоном, која је издвојена цртом од другог дела, тј. првим цифрама наших података, су представљена *стабла*, а у другом делу налазе се тзв. *листови*, тј. друге цифре наших података. Овај начин представљања података има своје предности: не губи се информација о појединачним подацима, а да се притом ипак постиже неки степен њихове компактификације.

Тачкасти дијаграми најчешће се формирају коришћењем статистичког софтвера. Погодан је за брзо утврђивање екстремалних вредности, а, такође, је користан када поредимо две или више серија података.

Пример 14. Одговарајући тачкасти дијаграм за претходни пример је