УНИВЕРЗИТЕТ У НИШУ
ЕЛЕКТРОНСКИ ФАКУЛТЕТ

- СЕМИНАРСКИ РАД -

КВАЛИТЕТ ПОДАТАКА

ПРЕДМЕТ: ПРИКУПЉАЊЕ И ПРЕДОБРАДА ПОДАТАКА
ЗА МАШИНСКО УЧЕЊЕ

Ментор: Александар Станимировић **Студент:** Филип Трајковић 1574

Ниш, 2023.

Садржај

1. Увод 3
2. Квалитет података 4
   2.1. Појам квалитета података 4
   2.2. Мере квалитета података 5
      2.2.1. Тачност 5
      2.2.2. Јединственост 6
      2.2.3. Конзистентност 6
      2.2.4. Потпуност 6
      2.2.5. Релевантност 7
      2.2.6. Правовременост 7
3. Расподела података 7
   3.1. Типови расподеле података 8
      3.1.1. Симетрична расподела 8
      3.1.2. Бимодална расподела 9
      3.1.3. Асиметрична расподела 10
      3.1.4. Униформна расподела 12
   3.2. Мере централне тенденције 13
      3.2.1. Аритметичка средња вредност 14
      3.2.2. Медијана 14
      3.2.3. Модус 15
   3.3. Врсте расподеле података 18
      3.3.1. Дискретна расподела података 23
      3.3.2. Континуална расподела података 28
4. Корелација 32
5. Варијанса 34
6. Примери из практичног дела семинарског рада 36
7. Закључак 39
8. Литература 40

1. Увод

Развој модерних софтверских и хадрверских система у савременом рачунарском окружењу утицао је на појаву све већег обима података. Велике количине података представљају основни садржајни материјал који се користи како би модерни софтверски системи имали висок ниво корисности у својим независним доменима примене. Како подаци представљају основни садржај сваког система у рачунарском окружењу, неопходно је над њима извршити низ корака који ће позитивно утицати на побољшање квалитативног нивоа тог садржаја. Низ корака који се спроводе у процесу побољшања квалитативног нивоа података се врше са циљем како би се употребна вредност самих података унапредила. Употребна вредност сваког појединачног податка је важан аспект из разлога извођења даљих закључака или проналаска законитости сагледавајући читав скуп података.

Низ корака који се спроводе са циљем побољшања квалитативног нивоа скупа података се односе на процес прикупљања и предобраде података који се обавља пре саме фазе анализе података или доношења закључака. Поступак прукупљања података је процес у којем се најпре врши процес агрегације прикупљених појединачних инстанци узорака при чему се највише обраћа пажња да квалитет узорака буде висок, односно да добијене вредности узорака буду приближне опште-познатом опсегу вредности који је базиран на доменском знању. Прикупљене узорке који се налазе изван овог опсега је могуће укључити у коначан скуп података уз обавезне додатне анализе у процесу предобраде података.

Неки од корака у процесу прикупљања и предобраде су процеси агрегације узорака, различитих испитивања квалитативних својстава узорака или целокупног скупа података, испитивања постојања појединачних вредности унутар скупа, испитивање валиданости добијених вредности узорака, процес анализе података у домену дистрибуције података, испитивање концентрације узорака око одређених вредности мера централне тенденције, сагледавање међусобних удаљености појединачних узорака на нивоу целокупног скупа података, упоређивање зависности између различитих фичера који се у скупу података налазе и др.

Претходно поменути кораци у процесу самог прикупљања и предобраде представљају обавезну фазу у процесу целокупне обраде и анализе података јер са собом носе испитивања елементарних и суштинских својстава сваке велике количине података. На основу добијених резултата у овој обавезној фази, добијају се мета-подаци о анализираном скупу на основу којих се могу препознати даљи кораци који се требају додатно укључити или искључити у процес даље анализе. Уколико добијени мета-подаци не задовољавају критеријуме квалитета података којима се овај рад бави, посматрани скуп података ће представљати ризичну групу узорака и потребно је одбацити такав скуп. Критеријуми квалитета података које треба сваки скуп да задовољи су обрађени у даљем тексту овог рада.

2. Квалитет података

Квалитет података се односи на развој и имплементацију активности које примењују технике управљања квалитетом на податке како би се осигурало да подаци одговарају специфичним потребама организације у одређеном контексту.

Квалитет података игра важну улогу у областима које се баве активностима везаним за анализу и обраду података при чему се на основу тих обрађених података могу извести одређене законитости из којих се креирају модели попут модела машинског учења. У процесима креирања модела машинског учења неопходно је да подаци носе високо квалитативна својства како би се процесу учења пружиле што прецизније и поузданије информације које тај скуп носи са собом, а које нису очигледне. Квалитет података у машинском учењу се односе на способност података да што више одговарају реално стању и потребама одређеног проблема који се покушава решити моделом машинског учења.

Подаци за које се сматра да одговарају њиховој намени сматрају се подацима високог квалитета. Тако добијени подаци високог квалитета представљају адекватну групу узорака над којима се касније могу вршити анализе успешности алгоритама, проблема у току моделовања и др.

2.1. Појам квалитета података

Појам квалитета података представља квалитативну меру добијеног узорка чија се вредност у процесу процене квалитета проверава у односу на претходно стечено знање из датог домена. Вредност квалитета сваког појединачног узорка се уз помоћ мера квалитета проверава на основу чега се доноси коначни закључак о квалитативним својствима посматраног узорка.

Уколико прикупљени подаци спадају у групу података који нису високог квалитета, процес даље анализе се мора обавити под високом дозом опреза при чему се и добијени резултати морају узети са резервом јер анализирани скуп не представља адекватну релевантну комбинацију вредности узорака за обављање истраживања.

Примери проблема са квалитетом података укључују дуплиране податке, непотпуне податке, недоследне податке, нетачне податке, лоше дефинисане податке, лоше организоване податке и лошу безбедност података. Како би се поменути проблеми са квалитетом уочили у адекватном временском интервалу пре иницирања процеса анализе података, неопходно је најпре извршити испитивања над добијеним скупом података користећи одређене мере квалитета како би се добила што јаснија слика о нивоу квалитета података који се обрађују.

2.2. Мере квалитета података

Неке од основних мера квалитета података су:

- Тачност
- Јединственост
- Конзистентност
- Потпуност
- Релевантност
- Правовременост

Слика 1. Основне мере квалитета података

2.2.1. Тачност

Представља меру квалитета података која дефинише вредност одступања података од стварне или исправне вредности оригиналног податка. Тачност је валидан избор евалуације за проблеме класификације података који су добро избалансирани и без великог броја “outlier”-а. Тачност вредности података се мери тако што се верификује у односу на познати извор тачних информација. Ово мерење може бити сложено ако постоји више извора који садрже тачне информације. У таквим случајевима, потребно је изабрати ону која највише утиче на домен

проблема и израчунати степен усклађености сваке вредности података са извором. Уз помоћ матрице конфузије можемо да дефинишемо тачност следећом формулом:

$$A = (T_p + T_n) / (T_p + F_p + T_n + F_n)$$

где су елементи:

A - вредност тачности, T_p - "True positive" вредност, T_n - "True negative" вредност, F_p - "False positive" вредност, F_n - "False negative" вредност

2.2.2. Јединственост

Јединственост података се односи на квалитативну меру сваког појединачног податка у великом скупу различитих података. Јединственост представља особину података која се односи на сваку појединачну ставку у подацима где се већим квалитетом подразумева и већа количина јединствених података. Јединственост података јесте супротност мултипликативности података тј. дуплицирања појединачних записа (врста) у табели података. Мултипликативност доводи до повећања обима скупа података без уношења варијабилности или различитости унутар скупа. Приликом процеса препроцесирања података, неопходно је извршити редукцију скупа података избацивањем дупликата - мултиплицираних вредности појединачних врста у табели.

2.2.3. Конзистентност

Конзистентност података се односи на униформност података док се крећу кроз мреже и апликације. Исте вредности података ускладиштене на различитим локацијама не би требало да буду различите. Један од начина да се осигура конзистентност података је детекција аномалија, која се понекад назива и анализа "outlier"-а (ванредних вредности), која помаже да се идентификују неочекиване вредности или догађаји у скупу података. Конзистентност проверава да ли су вредности података ускладиштене за исти запис у различитим изворима без контрадикторности и да ли су потпуно исте – у смислу значења, као и структуре и формата.

2.2.4. Потпуност

Потпуност података се односи на број попуњених вредности унутар скупа података што доприноси целовитости или свеобухватности скупа података. Како би скуп података био потпун неопходно је да не постоје празнине или недостајуће информације које на тај начин скуп података чине тежим за обраду и извођење даљих закључака из података. Непотпуни подаци су често неупотребљиви, али се често и даље користе чак и са информацијама које недостају, што даље може довести до скупих грешака и лажних закључака. Непотпуни подаци су често резултат неуспешно прикупљених података. Неке од техника које се примењују како би се отклонили проблеми непотпуних података су: попуњавање празних поља одређеним вредностима или извацивање целе

врсте или колоне из скупа података. Врста технике која се примењује зависи од броја недостајућих података у односу на целокупан скуп.

2.2.5. Релевантност

Релевантност података се односи на степен значаја информација које се налазе у одређеном скупу података. Релевантност података показује корисност типа података који се прибавља, његову квалитативну вредност и степен искоришћености тог податка у процесу обраде и анализе скупа података. Уколико подаци не носе високу релевантну вредност унутар целокупног скупа података, могуће их је одстранити како не би потенцијално утицали на коначни закључак процеса анализе података. Такође, процес прикупљања података није нимало јефтина операција, па је стога неопходно обратити пажњу на релевантност података који се прикупљају како целокупни процес не би уносио додатне трошкове у току анализе.

2.2.6. Правовременост

Правовременост података представља меру квалитета података која се односи на доступност и ажурираност података у одређеном временском тренутку и као таква мера представља једну од најважнијих мера које утичу на исправност и валидност података који се обрађују. Доношење битних пословних одлука се често ослања на валидност и доступност података којима се располаже, те је стога више него неопходно да подаци над којима се врши анализа буду правовремено одржавани и ажурирани у базама података. Правовременост такође утиче и на тачност података у одрђеном временском тренутку. Уколико подаци нису правовремено промењени или обновљени, може доћи до доношења погрешних закључака и акција на основу неисправних података. Метрика која се односи на правовременост података се изражава у процентима података који се могу добити у одређеном временском оквиру као што су дани, недеље, месеци, итд.

3. Расподела података

Расподела података представља меру која описује начин дистрибуције података у односу на целокупан скуп података према вредности учесталости понављања података у распону од минималне до максималне вредности. Расподела података се представља графиком вредности при чему су на x-оси представљене све вредности одређеног „feature“-а, тј. променљиве од интереса која се посматра, док су на y-оси представљене учесталости тих вредности у целокупном скупу података.

Дистрибуција се односи на то како су подаци распоређени или груписани око одређених вредности или опсега. Груписање око одређене вредности носи са собом високо квалитативну меру која представља начин простирања података у односу на учесталост у добијеном целокупном скупу узорака. Како бисмо детаљно описали расподелу података користе се одређени типови расподела који додатно доприносе процесу закључивања о структуираности и мери густине, квалитативних и квантитативних карактеристика

посматраног скупа података. Ови типови расподела описују обрасце по којима се подаци распоређују у целокупном скупу података и могу допринети побољшању разумевања законитости које владају у посматраном домену који је тим скупом података описан. Облик дистрибуције квантитативних (нумеричких) података се може одредити при чему треба постојати логичан редослед или распоред вредности при чему се могу идентификовати ниске (минималне) и високе (максималне) вредности на графику, као и густина распореда најзаступљенијих вредности у односу на мање заступљене вредности или граничне вредности које треба додатно испитати („outlier“-и).

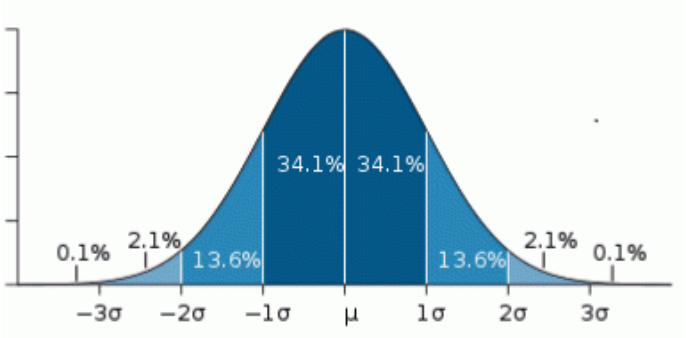
3.1. Типови расподеле података

Типови расподеле података представљају распоред варијабилности вредности података на нивоу целокупног скупа података. Уколико се подаци представе уз помоћ хистограма, типови расподеле података утичу на изглед графика тако што одређују где се налази највећи број вредности, око колико вредности су груписани сви узорци података, са којом тенденцијом раста или опадања се вредности скупа мењају и сл. Неки од основних типова расподеле података су:

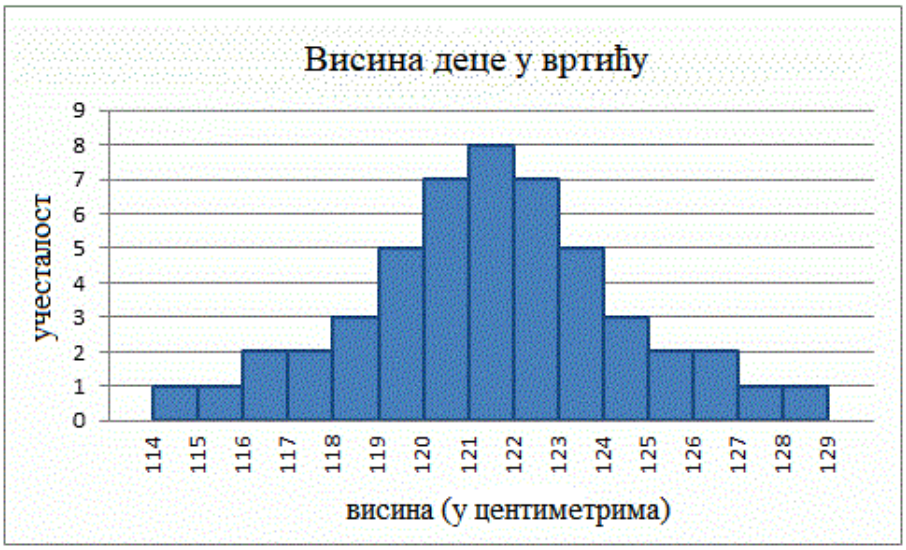
- симетрична расподела
- бимодална расподела
- асиметрична расподела
- униформна расподела

3.1.1. Симетрична расподела

Симетрична расподела представља расподелу података у којој две стране дистрибуције изгледају идентично у односу на вредност мере централне тенденције – средњу вредност. Нормална расподела представља идеалан пример за симетричну расподелу података. Изглед графика овакве расподеле је представљен у облику симетричног звона при чему се овај облик формира захваљујући облику колона које представљају податке. На следећем примеру можемо видети како симетрична расподела изгледа на хистограму:



Слика 2. Процентуална количина података у односу на средњу вредности - μ

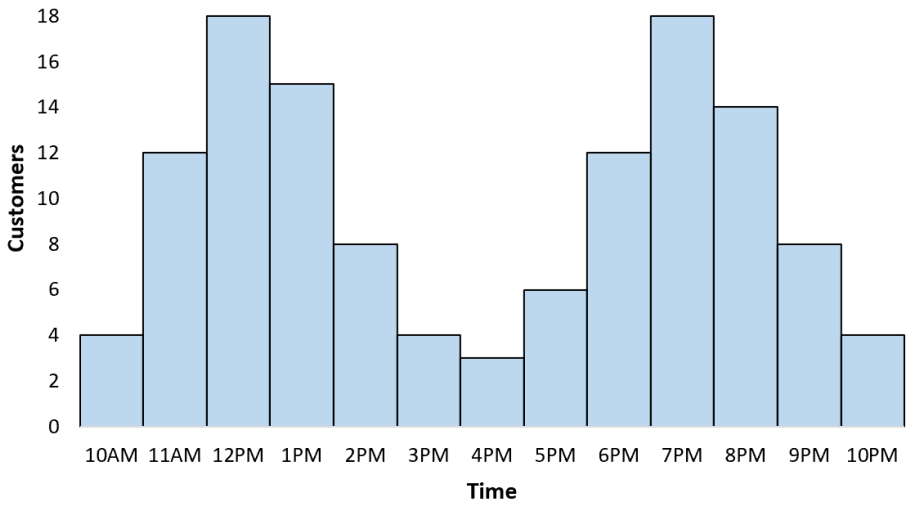


Слика 3. Симетрична расподела на примеру ученика

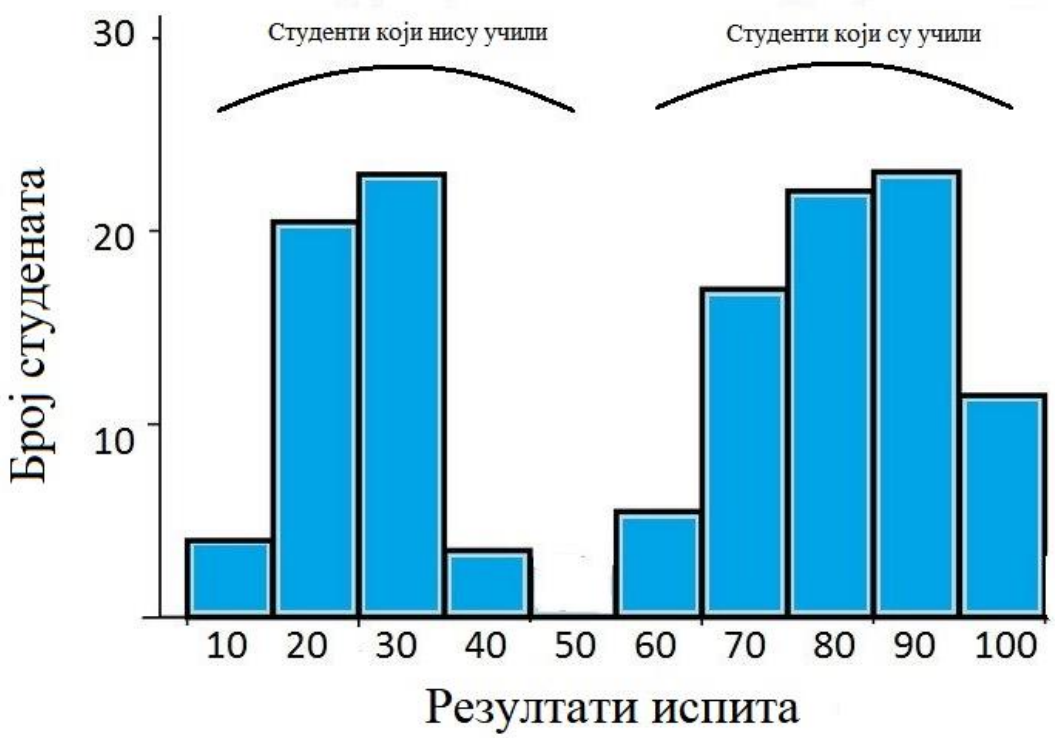
3.1.2. Бимодална расподела

Бимодална расподела података представља расподелу података чије се вредности концентришу око два најучесталија опсега вредности. Ова два опсега вредности учествују равноправно у расподели података те се стога све вредности из скупа података више или мање дисјунктно придружују једном или другом опсегу. Дисјунктно придруживање опсегу се односи на чињеницу да једна вредност не може истовремено припадати у оба опсега равноправно.

Оваква расподела података указује да у целокупан скуп података поседује две групе вредности што је на графику представљено са два „врха“. Око ових „врхова“ се гомилају све остале вредности. Бимодална расподела се често може представити и као две унимодалне расподеле које независно представљају своју групу узорака из скупа података.



Слика 4. Бимодална расподела на примеру посећености купаца у времену



Слика 5. Бимодална расподела на примеру резултата испита у односу на број студената

3.1.3. Асиметрична расподела

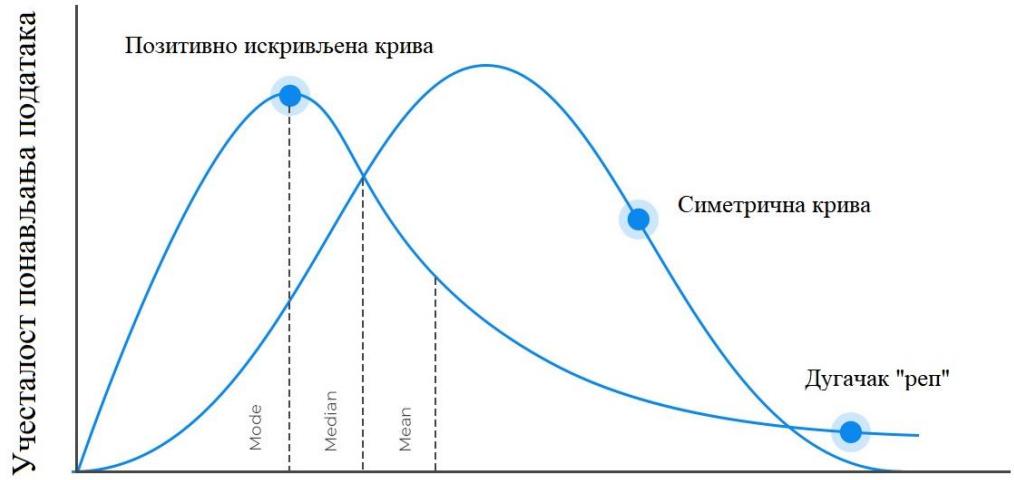
Асиметрична расподела представља тип расподеле података при чему график криве која описује податке није симетричан у односу мере централне тенденције скупа података. Овакав график има дугачки „реп“, односно вредност функције која умерено опада искључиво само са једне стране у односу на мере централне тенденције. Овакав „реп“ се креира на основу волумена података који се налазе изван опсега са најучесталијим вредностима у подацима.

Овакви типови података садрже велики број података чије вредности одступају од најучесталијих вредности из главног опсега. Најудаљеније вредности од главног опсега се називају „outlier“-има и такве вредности се након детаљне анализе у великом броју случајева одбацују из скупа података као невалидне. Овакве вредности могу утицати на коначан исход доношења закључака из података те је стога веома важно и неопходно ове вредности обрадити на адекватан начин.

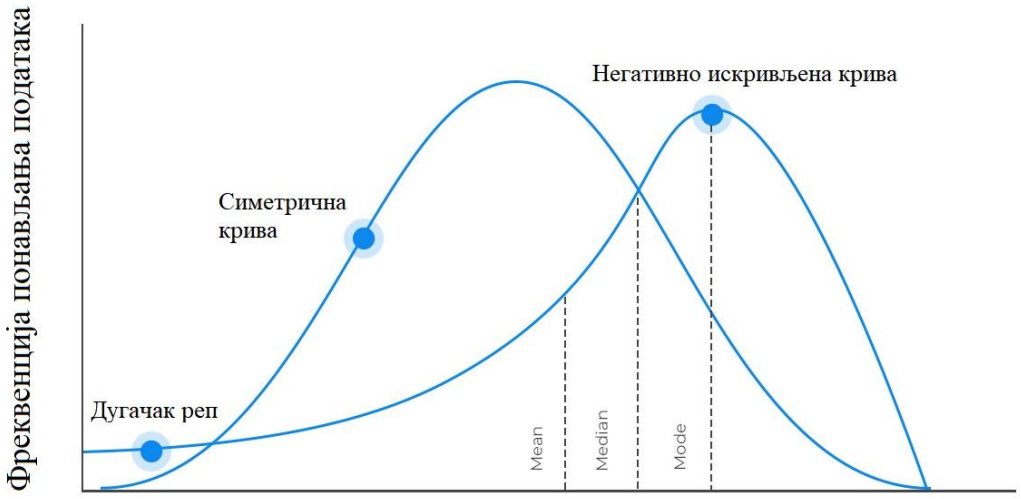
У зависности од тога са које стране функција учесталости спорије опада, разликујемо две врсте асиметричне расподеле података. Уколико функција спорије опада са десне стране у односу на опсег најучесталијих вредности, таква асиметрична расподела се назива – **позитивно** или **десно кошење** расподеле података. Уколико функција спорије опада са

леве стране у односу на опсег најучесталијих вредности, таква асиметрична расподела се назива – **негативно** или **лево кошење** расподеле података.

Битна карактеристика код асиметричне расподеле података јесте чињеница да су вредности мера централне тенденције различите једна од друге. Ова карактеристика је уочљива на следећим графицима који представљају горепоменуте асиметричне врсте расподеле података – позитивно и негативно кошење расподеле података.



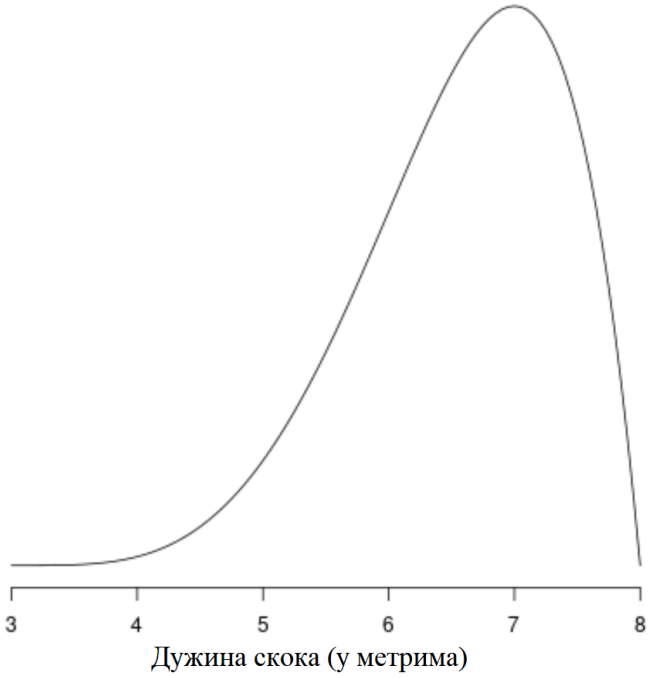
Слика 6. Позитивно или десно кошење расподеле података



Слика 6. Негативно или лево кошење расподеле података

На овим графицима може се јасно уочити дугачки реп који се образује на основу вредности изван опсега који садржи највећи број вредности и вредности „outlier“-а које представљају граничне вредности које се могу одбацити. Јасно се може уочити да се вредности „mean”, “median” и „mode“ разликују и то по вредности у зависности од типа кошења криве расподеле података.

Следећи пример показује реалан скуп података на којем су презентовани подаци из дисциплине скок у даљ, из спорта атлетике. На графику уочавамо лево кошење криве што означава да се у датом опсегу вредности неравномерно распоређују у односу на мере централне тенденције.



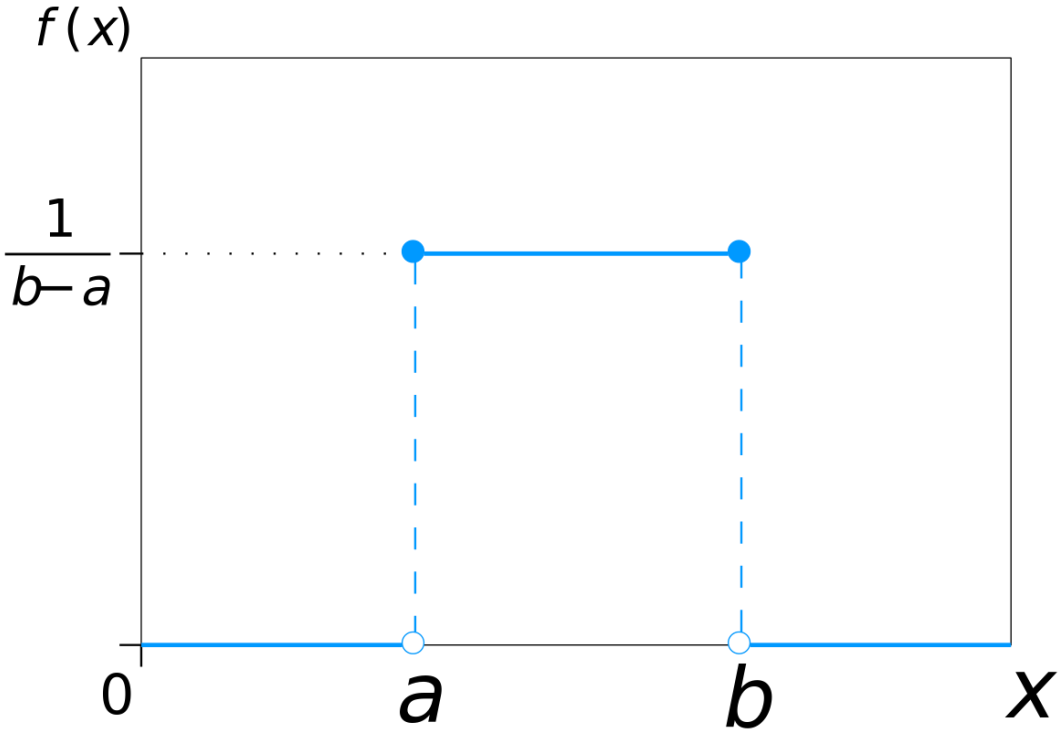
Слика 7. Пример левог кошења криве у атлетској дисциплини скок у даљ

3.1.4. Униформна расподела

Униформна расподела података је специјалан случај расподеле података при којој све вредности из скупа података са подједнаком учесталошћу учествују у расподели података.

Овај специјалан случај је представљен правом линијом између вредности a и b који чине интервал [a, b] свих могућих вредности унутар скупа при чему a представља минималну вредност, док b представља максималну вредност која се може добити. Вероватноћа P са којом се може добити свака појединачна вредност из посматраног скупа је:

$$P = \frac{1}{b - a}$$



Слика 7. График униформне расподеле података на опсегу [a, b]

3.2. Мере централне тенденције

Мере централне тенденције представљају фамилију вредности које имају за циљ да целокупан скуп података опишу уз помоћ једне јединствене вредности. Оваквим представљањем, тежи се да се скуп података представи неким заједничким својством.

Најбитнији елементи ове фамилије вредности су: аритметичка средња вредност, геометријска средња вредност, хармонијска средња вредност, средња квадратна вредност, средња кубна вредност, модус и медијана.

У даљем тексту ће бити презентовано неколико издвојених мера централне тенденције при чему свака носи другачију квалитативну вредност, али се свака од њих равноправно употребљава у односу на остале.

3.2.1. Аритметичка средња вредност

Аритметичка средња вредност представља најзаступљенију и највише коришћену меру централне тенденције. Аритметичка средња вредност узима у обзир све вредности из скупа података приликом процеса рачунања коначне вредности. Управо из разлога што узима све вредности из скупа, може довести до неправилних закључивања о скупу података зато што се у процесу израчунавања користе и граничне „outlier“ вредности које могу значајно утицати на коначну вредност израчунавања. Вредност аритметичке средине се израчунава као количник суме свих вредности из скупа података и броја узорака. Овај поступак је приказан на следећој формули:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ова мера централне тенденције има високу употребну вредност код скупова података који имају Гаусову (нормалну) расподелу података. Уколико је расподела података асиметрична тј. искошена неопходно је користити неки други приступ зато што тада средња вредност не приказује реалну вредност око које се подаци гомилају.

3.2.2. Медијана

Медијана представља меру централне тенденције која са вишим степеном веродостојности одређује средњу вредност у односу на аритметичку средњу вредност. У процесу израчунавања медијлне вредности се елиминишу вредности које могу негативно да утичу на реално стање расподеле података тако што елиминише граничне вредности – „outlier“-е.

Како би се израчунала медијална вредност, најпре је неопходно сортирани добијени вектор података у растући/опадајући редослед редослед. Уколико је број елемената вектора непаран, за медијалну вредност се узима елемент који се налази на средини датог низа.

Ово се може видети на следећем примеру.

Пример: Нека је дат низ непарног броја елемената са следећим бројевима:

65	55	89	56	35	14	56	55	87	45	92

Након сортирања овог низа бројева добија се следеће:

14	35	45	55	55	**56**	56	65	87	89	92

Обзиром да је дати низ непарног броја елемената, узимамо елемент на средини као коначну вредност медијане – 56.

Пример 2: Нека је дат низ парног броја елемената са следећим бројевима:

65	55	89	56	35	14	56	55	87	45

Најпре се врши корак сортирања улазних вредности и добија следеће:

14	35	45	55	**55**	**56**	56	65	87	89

Након сортирања, издвајају се два елемента са средине низа и врши се рачунање аритметичке средине између ова два броја.

$$Median = \frac{56 + 55}{2} = 55.5$$

Коначно решење овог примера је 55.5.

3.2.3. Модус

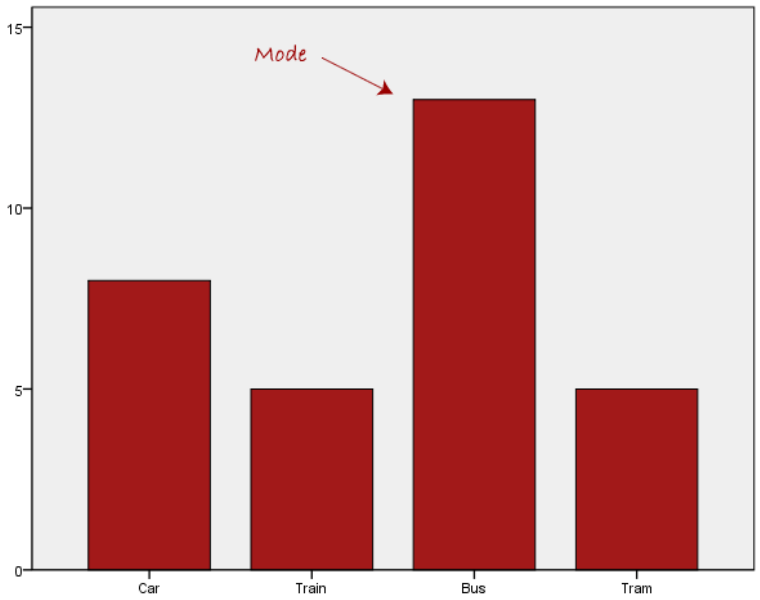
Модус представља меру централне тенденције при чему коначну вредност представља један или више узорака из посматраног скупа података. Модус вредност представља вредност узорка који има највећу вредност учесталости у скупу података. Уколико постоји више узорака који имају идентичну вредност фреквенције у скупу података, сви узорци се узимају као валидне модус вредности.

Ова мера централне тенденције се најчешће користи код фичера који могу имати мањи број могућих вредности као што су категорички подаци. Са повећањем броја могућих вредности које одређени фичер може да садржи, драстично опада квалитативна вредност коју ова мера носи са собом. Овај проблем се јавља из разлога зато што је код нумеричких података мања вероватноћа да ће вредности фичера имати потпуно идентичну вредност, јер се код мерења дешава одређен степен одступања. Због овог проблема, применљивост модуса је ниска код нумеричких типова података.

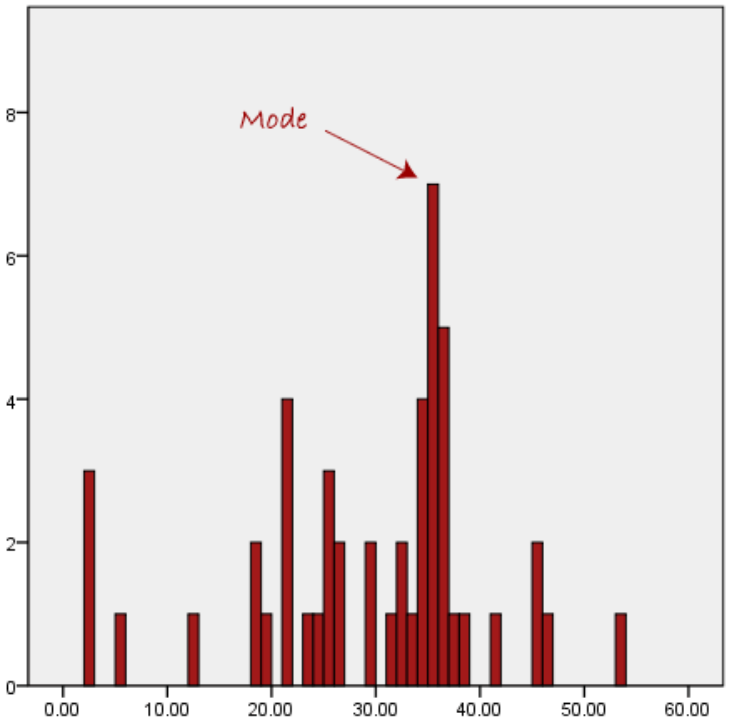
Код категоричких података код којих је број могућих вредности фичера мањи, ова мера може имати високо квалитативну вредност и може бити употребљена у различите сврхе.

Такође, још један проблем који се јавља јесте да је могуће да вредност модуса буде драстично удаљена од главног опсега у којем се налазе остале вредности из посматраног скупа података, па стога вредност модуса неће носити валидну информацију о целокупном скупу података.

Ова модус вредност је на хистограму представљена усправним правоугаоником са највећом висином, тј. највећом вредношћу на у-оси.

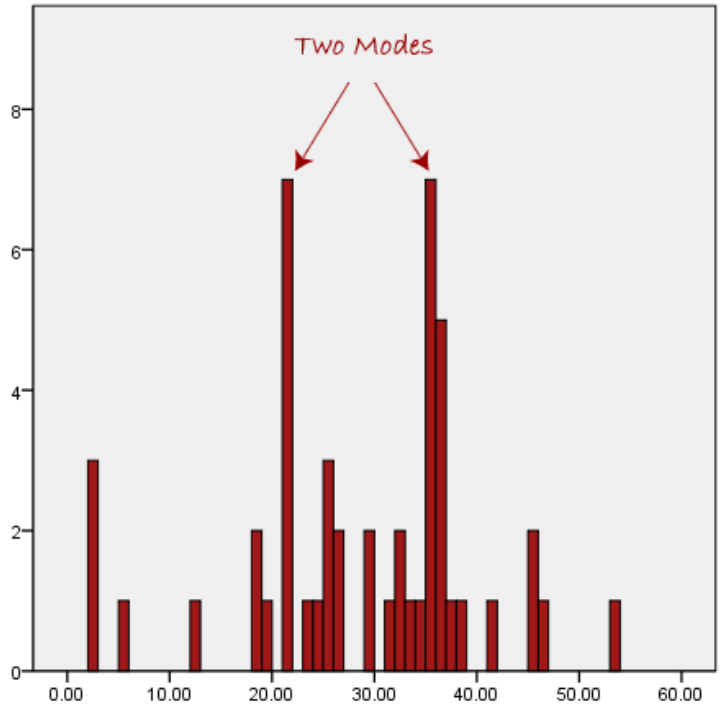


Слика 8. Модус категоричких података

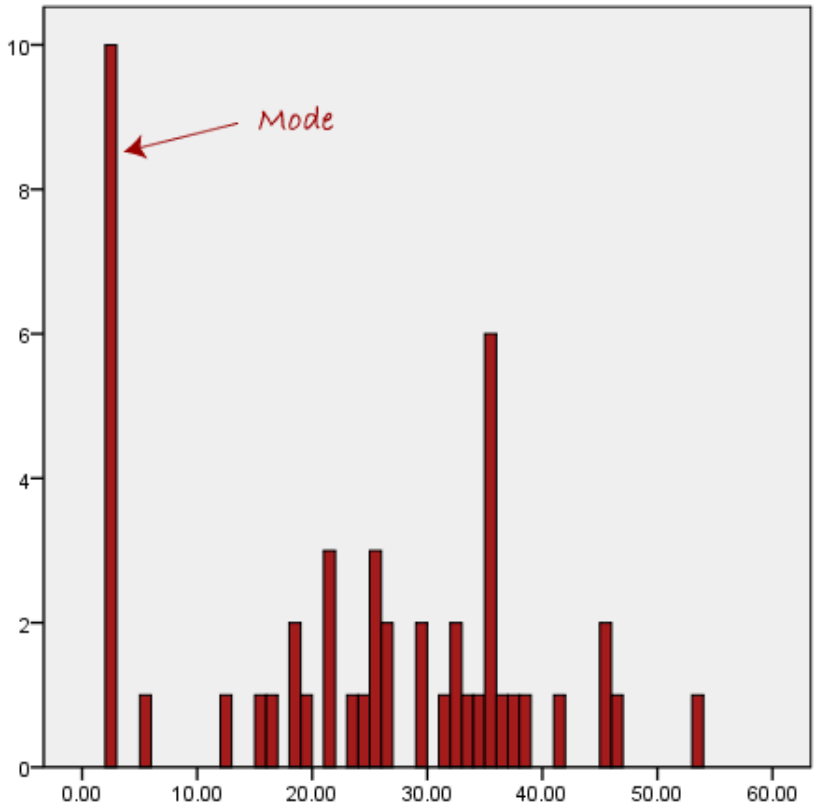


Слика 9. Модус нумеричких података

На следећим сликама су представљени неки од проблема који се јављају приликом израчунавања модуса.



Слика 10. Две вредности модуса са истом учесталошћу



Слика 11. Вредност модуса представљена „outlier”-ом

3.3. Врсте расподеле података

У зависности од типова података који се обрађују, расподелу података је могуће поделити у две репрезентативне врсте. Типови података који се могу презентовати одређеном дистрибуцијом су:

- дискретни подаци
- континуални подаци

Дискретни подаци представљају коначан скуп вредности који се може добити из одређеног улазног фичера при чему свака вредност представља бројиву карактеристику датог домена. Сваки дискретни коначан скуп вредности има ограничен број могућих нумеричких вредности којим се може квантификовати одређена акција или посматрани домен. Вредности из коначног скупа су представљени јединственим целим бројевима.

Континуални подаци представљају коначан скуп вредности у одређеном опсегу без јасне прецизности коју те вредности требају да садрже. Скуп вредности може имати бесконачан број могућих вредности и због тога је те вредности немогуће дискретизовати. Ове вредности могу бити било који реалан број из датог опсега.

Обзиром да оба типа података носе различите нумеричке вредности којима се може представити скуп података, врсте расподеле података делимо такође у две групе при чему ће за сваку групу бити обрађено неколико типичних представника:

- Дискретна расподела података
    1. Бернулијева расподела
    2. Биномиална расподела
    3. Поасонова расподела
- Континуална расподела података
    1. Нормална/Гаусова расподела
    2. „Log-Normal” расподела

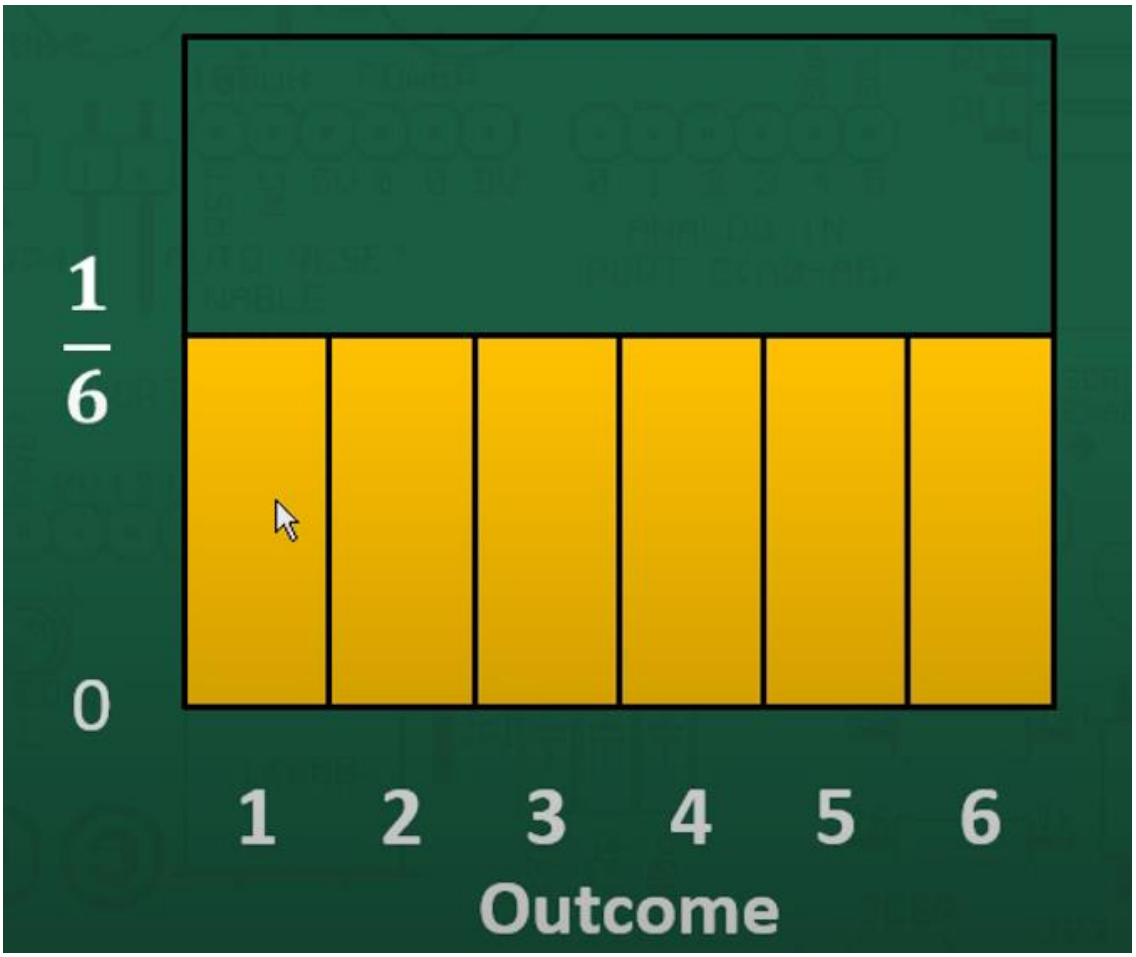
Како би разумевање различитих дистрибуција било јасније, неопходно је најпре разумевање типова графика којима ће поменуте расподеле бити приказане.

Дискретна расподела података се може приказати уз помоћ „PMF (eng. Probability Mass Function)" и „CDF (eng. Cumulative Distribution Function)" графика.

На следећем примеру је приказано како изгледају поменути графици приликом дитрибуције података код бацања коцке која има шест могућих вредности из скупа

[1, 6].

PMF – Probability Mass Function (код дискретне расподеле)



Слика 12. PMF график вероватноће код бацања коцке

На овом графику је на X-оси представљен могући потенцијални излаз (једна од шест могућности) са вредностима од 1 до 6. На У-оси је приказана заједничка вредност вероватноће која важи за сваку од могућих вредности.

Главна карактеристика овог типа графика је да је скуп коначних вредности приказан усправних правоугаоницима за сваку појединачну могућност.

Оваква униформна расподела није карактеристика дискретне расподеле података, већ овог приказаног примера.

CDF – Cumulative Distribution Function (код дискретне расподеле)

Слика 13. CDF график вероватноће код бацања коцке

CDF график представља вероватноћу за свако од бацања при чему се за сваку вредност узима и сума вредности вероватноћа за могућности које имају мању вредност од посматране. Следеће формуле демонстрирају начин израчунавања вероватноће за сваку од могућих вредности овог примера.

$$P(X \leq 1) = P(X = 1)$$

$$P(X \leq 2) = P(X = 1) + P(X = 2)$$

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3)$$

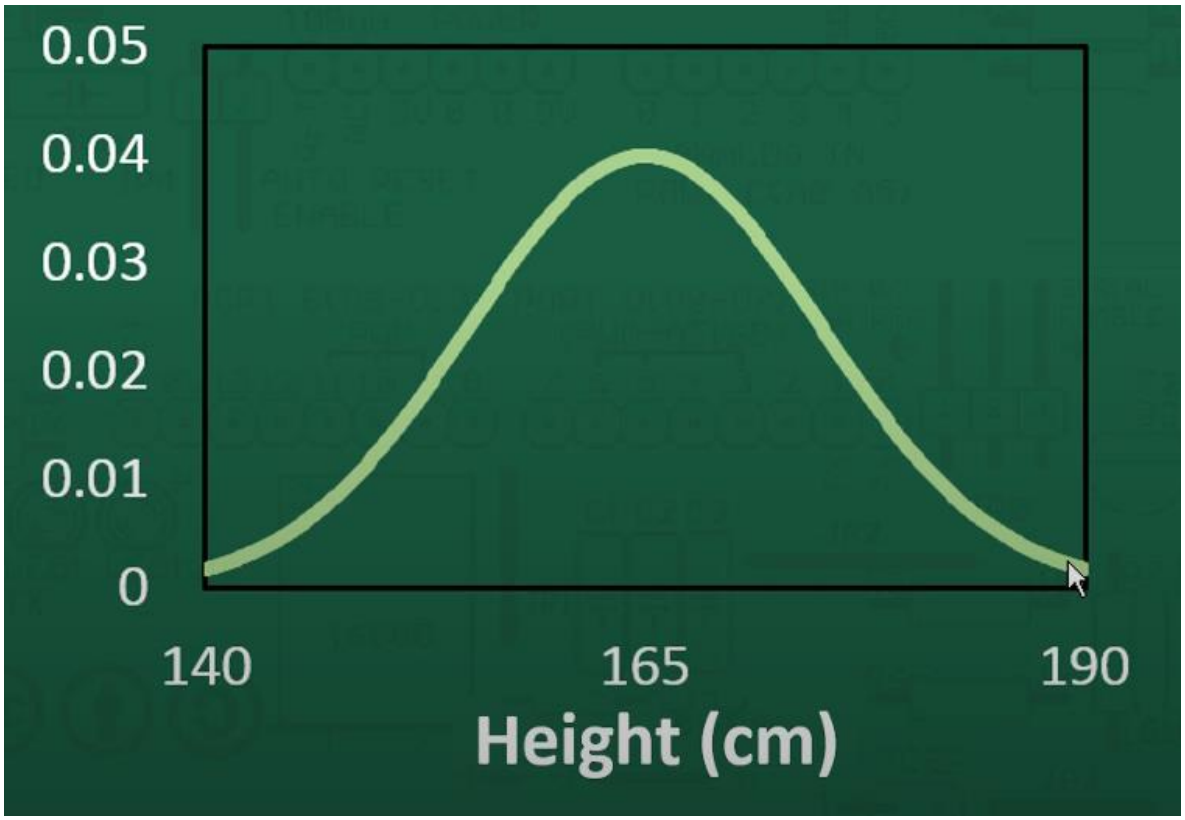
$$P(X \leq 4) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$P(X \leq 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

$$P(X \leq 6) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$

Континуална расподела података се може приказати уз помоћ „PDF (eng. Probability Density Function)" и „CDF (eng. Cumulative Distribution Function)" графика.

PDF – Probability Density Function (код континуалне расподеле)

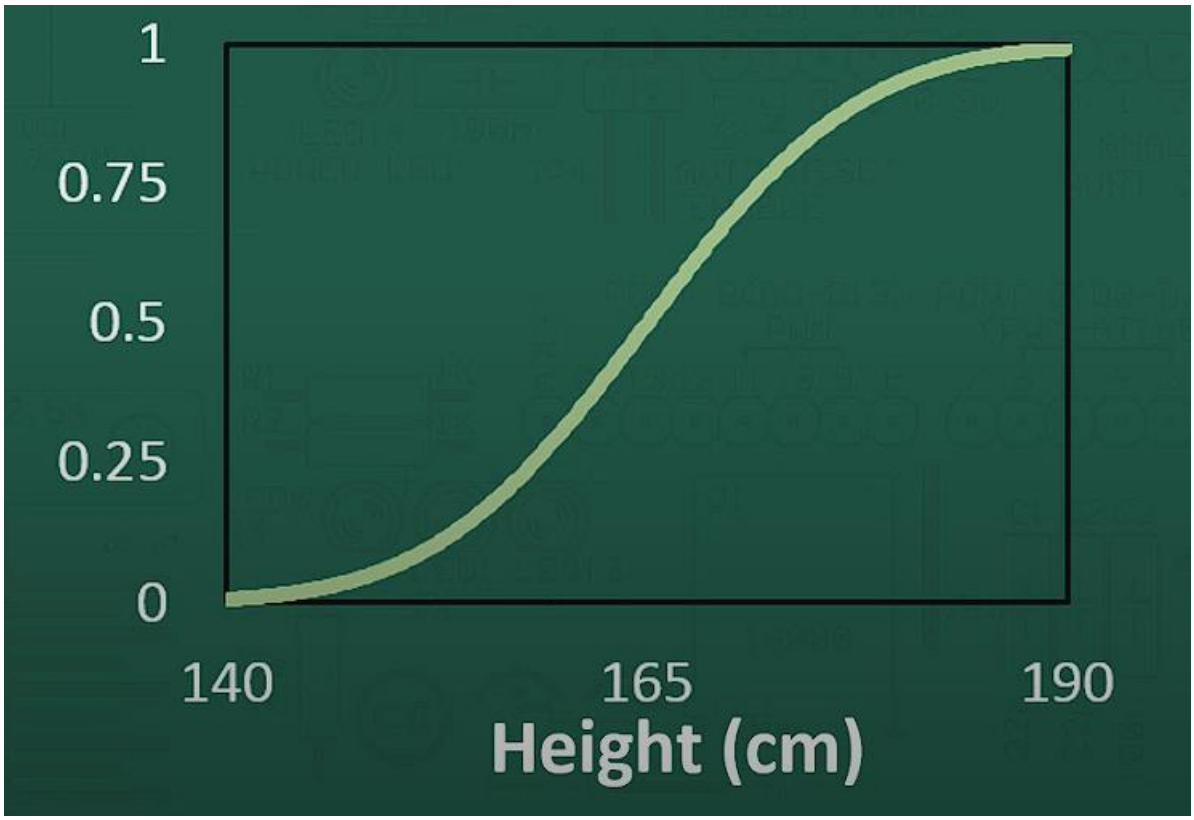


Слика 14. PDF график вероватноће код висине испитаника

На овом графику се примећује нормална/Гаусова расподела при чему се вредности узорака крећу у опсегу између 140 и 190 центиметара при чему је вредност 165 цм мера централне тенденције око које се гомила највећи удео вредности свих узорака. На У-оси су представљене вероватноће појављивања појединачних узорака у целокупном скупу података.

Приликом формирања графика овог типа, јасно је уочљиво да је број могућих вредности веома велики тј. да је скуп података континуалног типа. На графику су вредности вероватноће представљене кривом вероватноћа која повезује вероватноће свих узорака у посматраном скупу података.

CDF – Cumulative Distribution Function (код континуалне расподеле)



Слика 15. CDF график вероватноће код висине испитаника

На овом графику је приказ идентичан пример расподеле вероватноће висине испитаника као на претходном графику. На X-оси су вредности опсега посматраног скупа података што се идентично поклапа као код PDF приказа. На У-оси је приказан удео вредности или проценат целокупног скупа података од почетка опсега до дате вредности са X-осе.

Вредност нагиба криве је директно пропорционална вредности вероватноће са PDF приказа. Тачније што је вредност вероватноће на PDF графику већа, то ће код исте вредности на CDF графика нагиб криве бити већи. Обзиром да је вредност вероватноће највећа код средње вредност скупа 165 цм, такође је и нагиб криве највећи код ове вредности док се уједначено нагиб криве смањује како се вредност функције приближава граничним вредностима посматраног скупа. Уколико вредност функције на PDF графику представимо са f(x), а вредност функције на CDF графику са F(x) можемо извести следеће формуле уз помоћ којих је могуће приказати један график на основу другог и обрнуто.

$$\frac{dF(x)}{dx} = f(x)$$

$$\int_{-\infty}^{x} f(x)dx = F(x)$$

3.3.1. Дискретна расподела података

У овом поглављу ће бити приказани главни представници дискретне расподеле података.

3.3.1.1. Бернулијева расподела

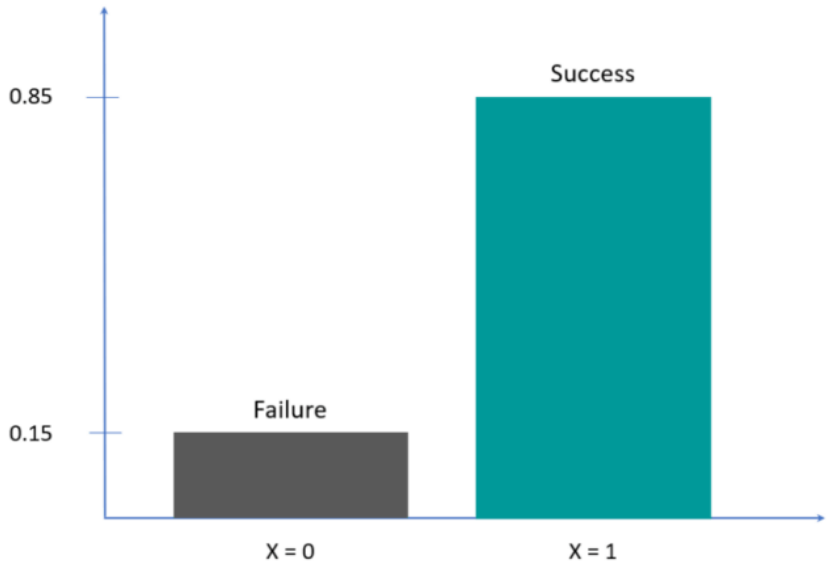
Бернулијева расподела представља најједноставнији облик дискретне расподеле података при којој се расподела вероватноће врши за тачно две могућности које се могу десити. Помоћу ове расподеле се моделује однос две могућности попут бацања новчића, бирања између две вредности „тачно“ или „нетачно“ и сл. Вероватноћа која се добија на овом графику представља вероватноћу акције која се само једном извршава што укључује једно бацање новчића или један одговор на питање са тачним и нетачним могућностима. На основу вероватноће једне могућности, може се одредити вероватноћа друге могућности према следећем правилу. Уколико је вероватноћа једне могућности означена са $p1 \in [0, 1]$, онда је вероватноћа друге могућности p2 једнака:

$$p2 = 1 - p1$$

Комплетна формула за све могуће случајеве вредности могућих излаза јесте:

$$\mathbf{P(X = x) = p^x(1 - p)^{1-x}; P(x) = \begin{cases} 1 - p, & \text{for } x = 0 \\ p, & \text{for } x = 1 \end{cases}}$$

где је x – могућа вредност излаза (0 или 1), а p – вероватноћа успеха.



Слика 16. Бернулијева расподела за случајеве успеха (p = 0.85) и неуспеха

3.3.1.2. Биномиална расподела

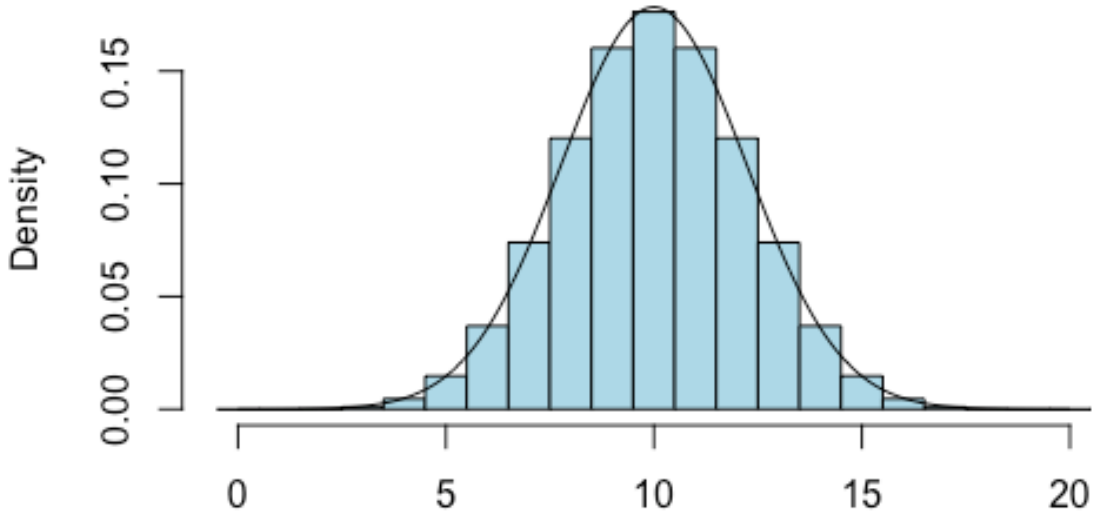
Биномиална расподела представља дискретну расподелу података која сумира вредности више извршених итерација акција као што су акције код Бернулијеве расподеле. Главни предуслови итерација јесу да свака итерација може имати искључиво два излаза (успешно или неуспешно), да је вероватноћа успеха или неуспеха идентична приликом извнођења свих итерација и да су све итерације које се извршавају независне једна од друге.

Формула вероватноће код биномиалне расподеле се рачуна на следећи начин:

$$P\left(X = x\right) = \frac{n!}{x!(n-x)!}\, p^x \left(1 - p\right)^{(n-x)}$$

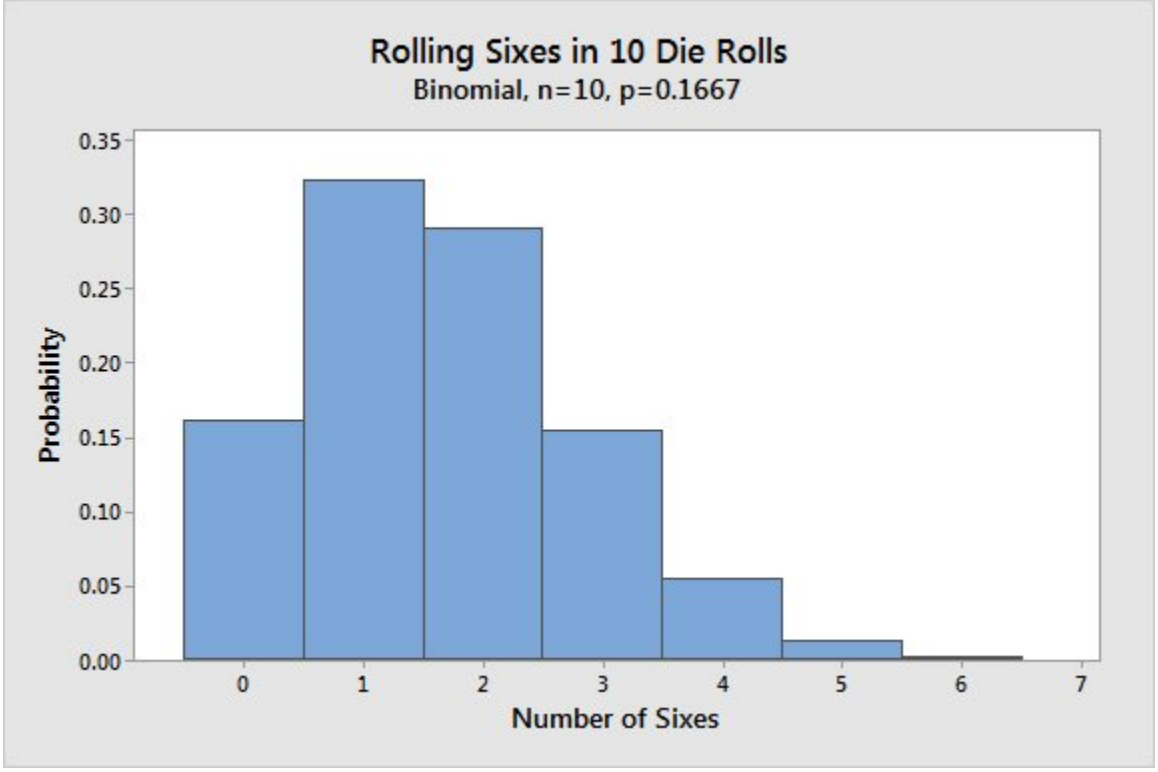
где је n – број итерација које се обављају, а p – вероватноћа да ће итерација бити успешна

На следећем графику је приказана сличност између графика биномиалне расподеле (са плавим правоугаоницима) и нормалне расподеле (представљене црном континуалном линијом). За сваку вредност је одређена вредност густине тј. учесталошћу у појављивању у односу на целокупан скуп података. Уочавамо симетричну расподелу на овом примеру и код биномиалне и код нормалне расподеле.



Слика 17. Пример биномиалне расподеле

На следећем примеру је приказана биномиална расподела вероватноћа код бацања коцке у десет итерација при чему се рачуна вероватноћа добијања броја 6 у свакој од итерација. Вероватноћа добијања броја шест у свакој итерацији је једнака p = 0.1667.



Слика 18. Биномиална расподела за случај бацања коцке 6 пута са циљем добијања броја шест

На графику је јасно уочљиво да се вредност вероватноће добијања већег броја шестица смањује, а да је вероватноћа добијања једне шестица највећа.

3.3.1.3. Поасонова расподела

Поасонова расподела представља дискретну расподелу података која се бави вероватноћом да ли ће се или неће одређени догађај догодити у одређеном интервалу који се посматра. Како би се израчунала вероватноћа догађаја одређене ситуације у неком интервалу, неопходно је претходно познавање вероватноће понављања тог догађаја у одређеном приближном интервалу.

Главне карактеристике или предуслови:

- да су догађаји који се посматрају независни једни од других
- да се догађај може десити неограничен број пута

- да се два догађаја не могу десити истовремено

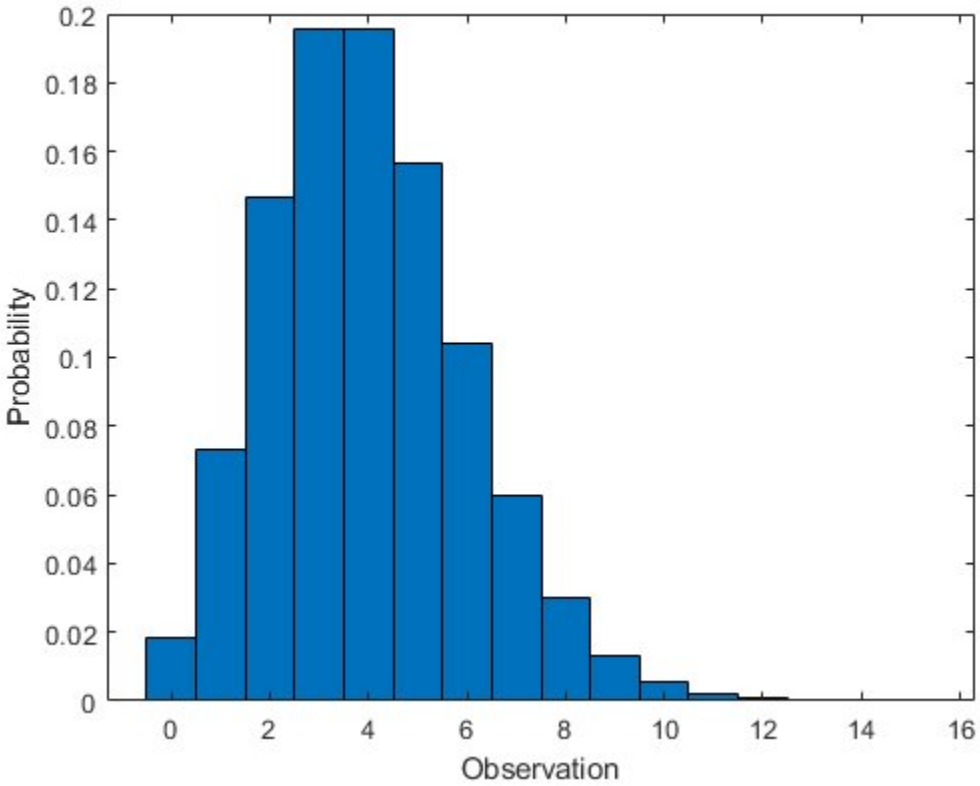
Формула поасонове дистрибуције је дата на следећи начин:

$$f(x) = \frac{\lambda^x}{x!}e^{-\lambda}$$

где је λ – средња вредност броја понављања посматраног догађаја, x – број понављања за који се одређује вероватноћа, e – Ојлерова константа

На следећем графику је представљена вероватноћа учесталости одређеног броја догађаја у истом временском интервалу.

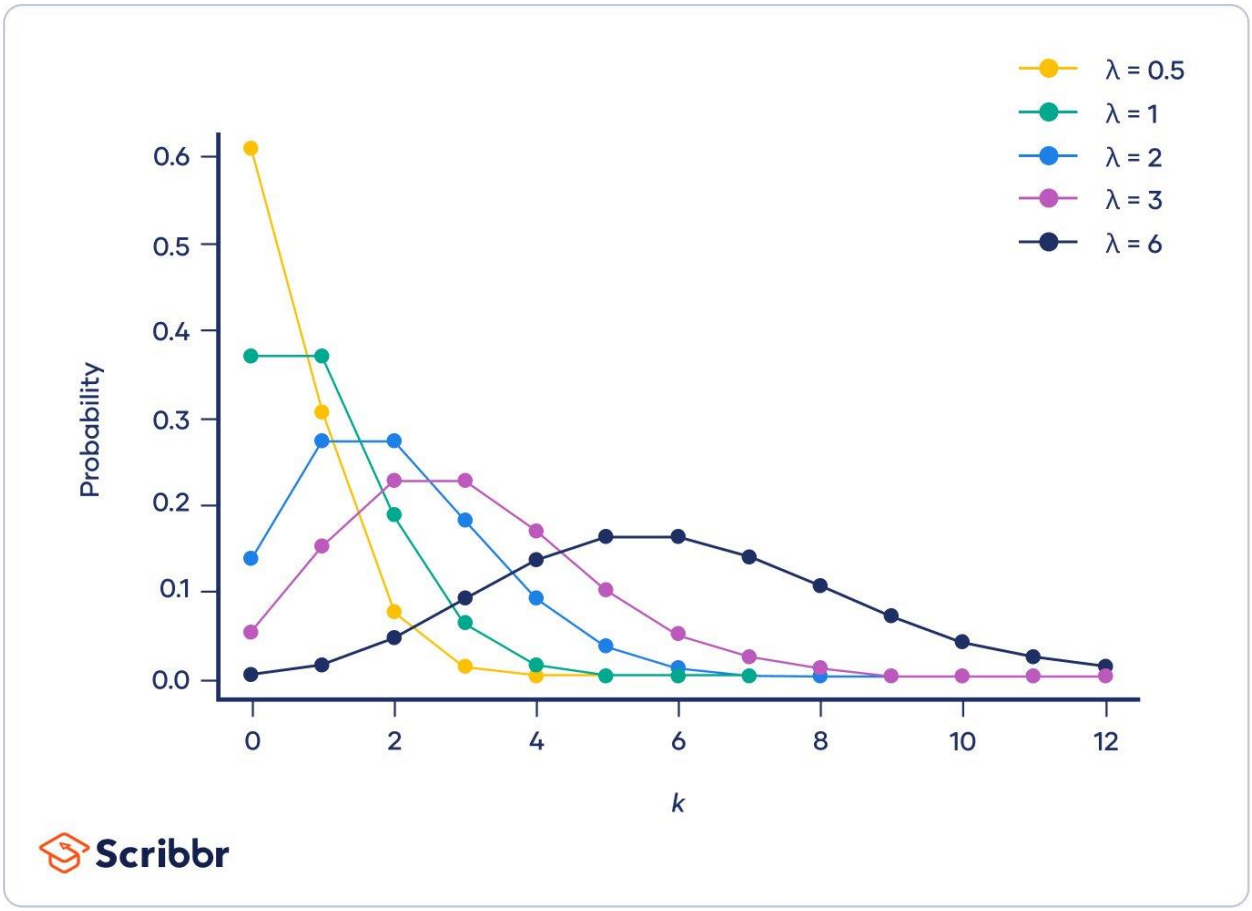
Са графика је јасно уочљиво да се посматрани догађај у одређеном интервалу најчешће дешава три или четири пута.



Слика 19. График Поасонове дистрибуције

Битно је напоменути да су вредности варијансе и средње вредности Поасонове дистрибуције једнаке вредности λ која представља средњу вредност понављања одређеног догађаја у посматраном периоду.

Вредност λ има великог утицаја у одређивању вероватноће код Поасонове дистрибуције па је управо та условљеност приказана на следећем графику. На слици 20 приказане су различите вредности Поасонове дистрибуције код којих се разликује само λ вредност.



Слика 20. График Поасонове дистрибуције за различите вредности λ

Овај тип дистрибуције се користи када се тачно време посматраног догађаја у одређеном временском интервалу не може предвидети тако да се дешава са случајном вероватноћом понављања.

Неки од примера када је добро користити овај тип дистрибуције: одређивање броја купаца у продавници у одређеним временским интервалима, вероватноћа учесталости телефонских позива у посматрано време, вероватноћа постизања одређеног броја поена у спортским такмичењима на основу претходних утакмица и сл.

3.3.2. Континуална расподела података

Континуална расподела података представља функцију вероватноће понављања одређених континуалних вредности која се не могу груписати у оквиру одређеног интервала или се нормализовати на одређени цео број. Вредности које се налазе у посматраном скупу података су представљене реалним бројевима, а број различитих вредности које тај скуп може да садржи је веома велики.

У следећим поглављима су обрађени типични представници ове групе расподеле података.

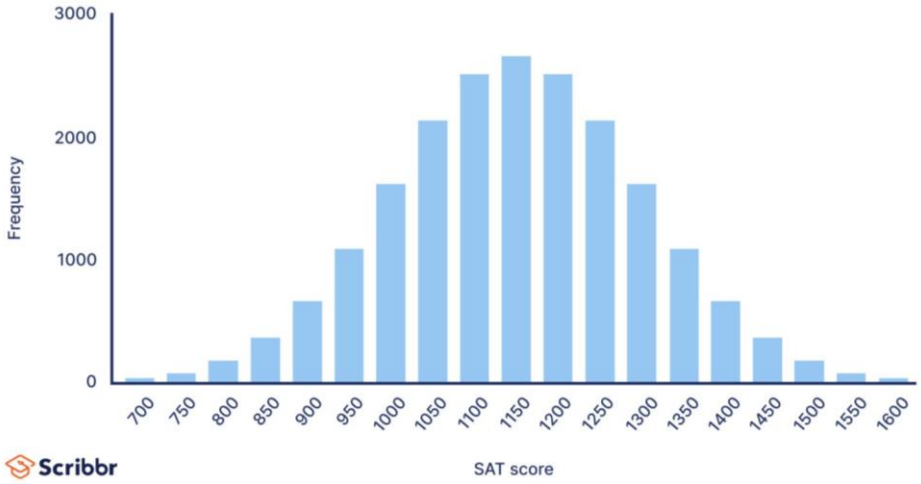
3.3.2.1. Нормална/Гаусова расподела

Нормална или Гаусова расподела је најједноставнији пример континуалне расподеле података која је представљена симетричном расподелом на графику. Крива функције ове расподеле има звонасти облик што указује да се ради управо о симетричној расподели.

Главна карактеристика података који су представљени овом расподелом јесте да су мере централне тенденције средња вредност, медијана и модус једнаке. Такође, број могућих вредности мањих од медијане је једнак броју могућих вредности већих од медијане целокупног скупа података. Уз помоћ ове расподеле, могуће је извршити предикцију са одређеном тачношћу вероватноће на основу претходних трендова, тј. промена у подацима.

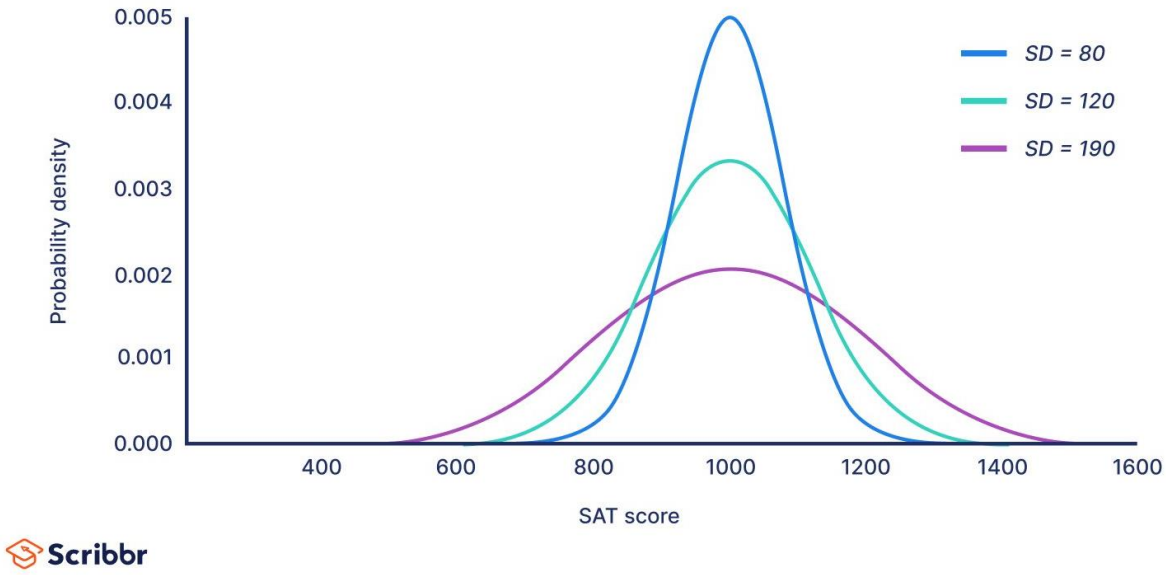
Обзиром да је дистрибуција симетричног облика, не постоји лево или десно кошење расподеле података.

Површина испод криве функције Гаусове расподеле података мора бити једнака јединици.



Слика 21. График Гаусове дистрибуције

На графику приказаном на слици 22 представљена је Гаусова расподела података при чему се разликују вредности стандардне девијације за сва три посматрана скупа података. Вредности стандардне девијације су приказане са десне стране.



Слика 22. Графици Гаусове дистрибуције са различитим вредностима стандардне девијације

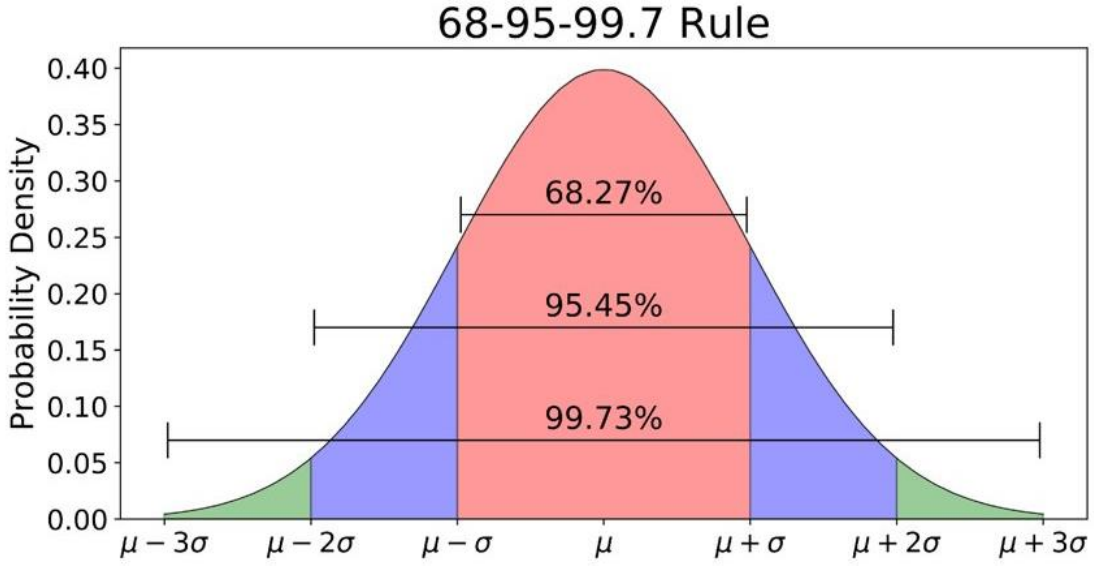
68-95-99.7 Правило

Ово правило представља емпиријску законитост изведену на основу различитих Гаусових дистрибуција података. Приликом представљања графика, 68% свих вредности скупа података се налазе у опсегу вредности једне стандардне девијације од средње вредности.

Такође, према овом правилу важи да се 95% вредности скупа података налази у опсегу вредности највише удаљеном за вредност две стандардне девијације од средње вредности скупа података.

Вредности које се налазе на удаљености максимално три стандардне девијације од средње вредности, налазе се у 99.7% свих вредности скупа података док се вредности изван тог опсега могу одбацити као „outlier“-и тј. невалидне вредности.

На следећем графику је приказано ово правило са одговарајућим опсезима вредности које у те опсеге спадају.



Слика 23. *68-95-99.7 правило* са означеним опсезима вредности

3.3.2.2. „Log-Normal" расподела

"Log-Normal" расподела података представља десно искривљену расподелу података са дугачким „репом" који се простира у десну страну. Вредност функције почиње из нулте вредности и стрмо расте до максималне вредности функције, након чега вредност функције постепено опада.

Вредност густине вероватноће функције зависи од два параметра μ и σ , при чему је вредност X > 0. Формула ове расподеле је представљена на следећи начин:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}$$

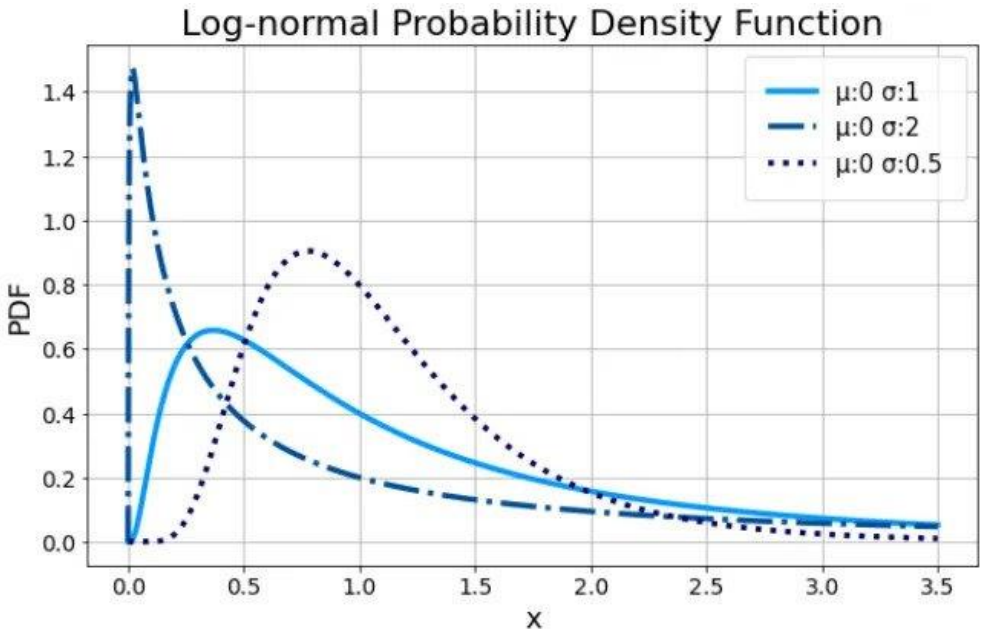
где је: μ - локациони параметар, а σ – параметар скалирања дистрибуције.

Ове две вредности је могуће израчунати на следећи начин.

$$\hat{\mu} = \frac{\sum_k ln\, x_k}{n} \text{ and } \hat{\sigma}^2 = \frac{\sum_k (\ln x_k - \hat{\mu})^2}{n}$$

Ова два параметра не представљају вредности стандардне девијације или средње вредности код овог типа дистрибуције. Средњу вредност и стандардну девијацију је могуће добити из ових вредности применом одређених логаритамских функција.

На следећем графику се могу видети различите криве графика за различите вредности параметра скалирања дистрибуције на „PDF" графику.

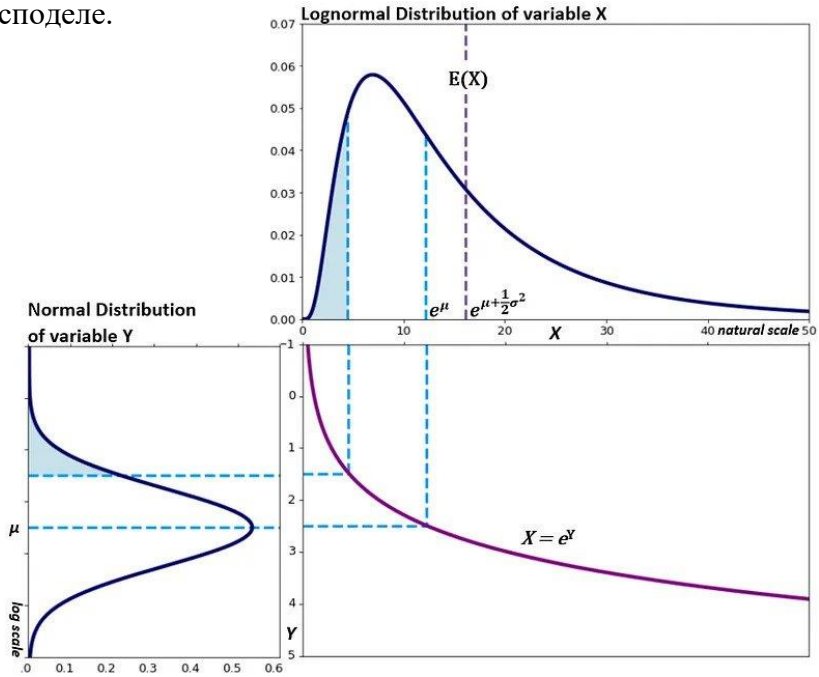


Слика 24. *„Log-Normal“ расподела* са различитим вредностима параметра скалирања

„Log-Normal“ расподелу је могуће претворити у Гаусову расподелу применом одређених логаритамских метода при чему се може одрадити пресликавање као на слици 25.

Када на вредности X из „Log-Normal“ расподеле пресликамо уз помоћ $Y=\ln(X)$ добијамо вредности Y које се распоређују према Гаусовој расподели.

Ова погодност трансформације омогућава нам да се подаци дистрибуирани према „Log-Normal” расподели прилагоде методама обраде података за Гаусову расподелу како би се обавио низ анализа података. Податке је затим могуће трансформисати у основни облик логаритмске расподеле.



Слика 25. Пресликавање „Log-Normal“ расподела у Гаусову расподелу

4. Корелација

Корелација представља нумеричку вредност линеарне зависности између појединачних фичера унутар посматраног скупа података. Приликом рачунања корелације, врши се упоређивање сваког фичера са сваким од преосталих фичера из скупа података при чему се за сваку везу добија зависност у одређеном нумеричком опсегу. Моделовањем зависности на овај начин, могу се изразити одређена имплицитна међусобна својства између података. Подаци попут међусобних зависности између појединачних фичера могу бити од суштинске користи зато што носе високу употребну и квалитативну вредност у процесу анализе података и закључивања законитости које делују у посматраном скупу података.

Вредност корелације или Пирсонов коефицијент корелације између два посматрана фичера се рачуна као количник вредности коваријансе између два фичера и производа појединачних вредности стандардних девијација поменутих фичера. На овај начин се вредност коваријансе између две фичера нормализује у опсегу од [-1, 1].

Вредност корелације може бити негативна или позитивна, што је условљено знаком испред вредности корелације.

Уколико је вредност корелације негативна, линеарна законитост која важи између два фичера је обрнуто линеарно зависна. Негативна корелација означава да се вредност једног фичера смањује уколико се вредност другог фичера повећава и обрнуто.

Уколико је вредност корелације позитивна, линеарна законитост која важи између два фичера је директно линеарно зависна што означава да се вредност једног фичера повећава са повећањем вредности другог фичера и обрнуто.

Уколико је вредност корелације између два фичера једнака нули, линеарна зависност између два фичера не постоји. Вредности корелације које су позитивне или негативне вредности приближне нули представљају слабо условљену корелацију између фичера и не представљају довољно квалитативну меру на основу које се може утврдити висока зависности између фичера.

Вредности корелације које су у опсегу [-1, -0.5] или [0.5, 1] представљају високо условљену зависност и такву линеарну релацију између посматраних фичера треба јасно издвојити јер се на основу те релације могу извести одређени високо квалитативни закључци о скупу података.

Приликом израчунавања вредности корелације на нивоу целокупног скупа података, применом уграђених функција из пајтон библиотека, добија се матрица корелације – симетрична матрица која за врсте и колоне има улазне фичере посматраног скупа података. На главној дијагонали добијене матрице се налази вредност 1 јер је главна дијагонала пресек врста и колона истог фичера. У остала поља ове матрице се смештају

вредности корелације између свих осталих фичера појединачно и то у пресеку врста и колона свих фичера понаособ.

Формула рачунања коваријансе два улазна фичера X и Y:

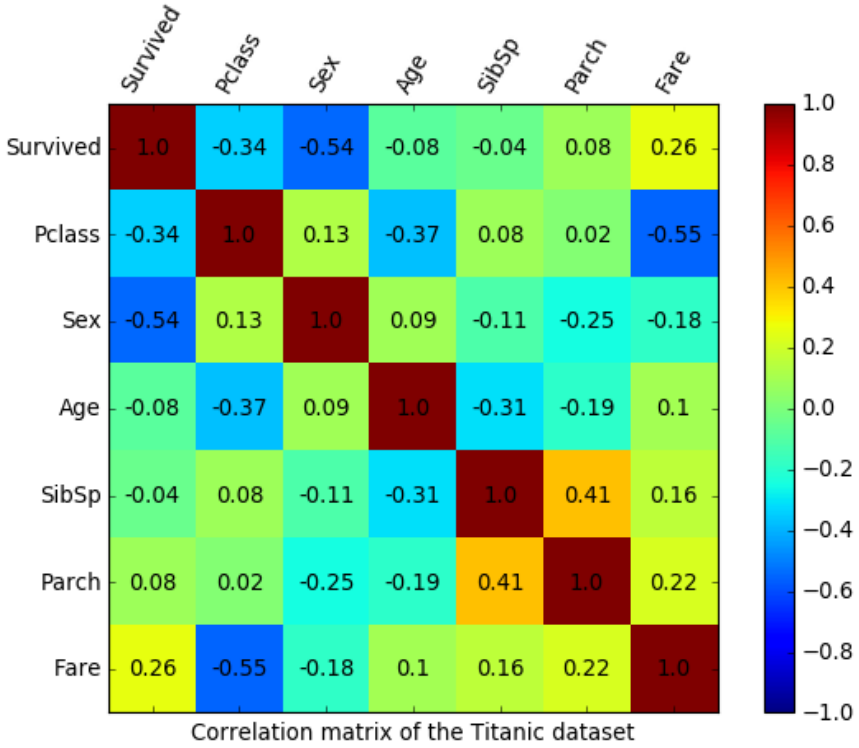
$$cov(x,y) = \frac{1}{n} \times \sum_{i=1}^{n}(x_i - mean(x)) \times (y_i - mean(y))$$

где су: X, Y – улазни фичери, n – број елемената улазних фичера X и Y

Након израчунавања коваријансе улазних фичера, могуће је извршити рачунања вредности Пирсоновог коефицијента корелације на следећи начин:

$$\text{Pearson's correlation coefficient} = \frac{cov(x,y)}{stdev(x) \times stdev(y)}$$

На следећем примеру је приказана матрица корелације за скуп података о путницима Титаника са различитим вредностима корелације између појединих фичера. Градијенталним опсегом боја је представљен степен корелације између фичера што се може видети са десне стране слике.

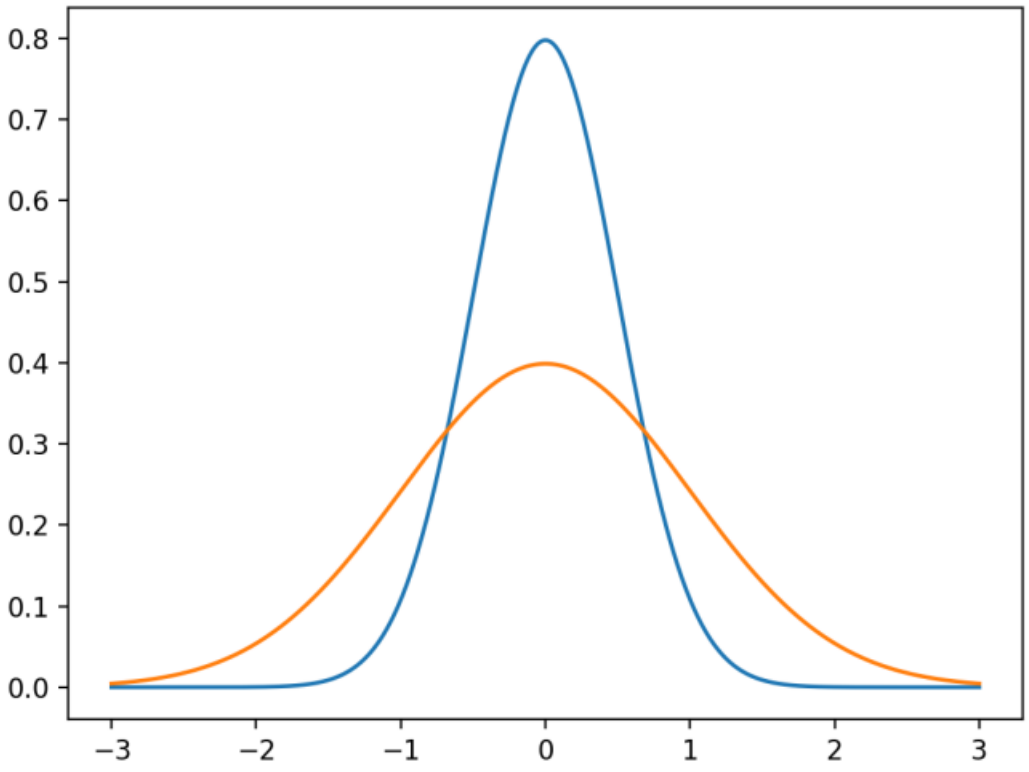


Слика 26. Матрица корелације скупа података о путницима Титаника

5. Варијанса

Варијанса представља меру која показује колико просечна вредност узорака одступа или варира од мере централне тенденције – средње вредности. Ова мера детаљније описује начин ширења вредности у целокупном скупу података. Варијанса може имати ниску или високу вредност што утиче на то колико ће подаци бити груписани око средње вредности. Скуп података са ниском варијансом ће имати вредности груписане у близини средње вредности, што значи да ће нагиб криве одређен расподеле података бити већи око средње вредности. Скуп података са високом варијансом ће имати вредности које су удаљеније од средње вредности и, на примеру Гаусове расподеле података, представљаће облик широког звона.

Вредност варијансе не може бити негативна, тачније налази се у скупу вредности [0, +∞).



Слика 27: Изглед Гаусове дистрибуције података са високом(плава крива) и ниском(наранџаста крива) варијансом

Варијанса може имати два различита облика:

- Варијансу узорака
- Варијансу популације

Варијанса узорака

Варијанса узорака Гаусове расподеле података се рачуна као средња квадратна вредност разлике сваког појединачног узорка из скупа у односу на средњу вредност целокупног скупа података.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Слика 28: Формула варијансе узорака

где је: s^2 – варијанса узорака, x_i – вредност i-тог елемента узорка, $\bar{x}$ - средња вредност свих узорака, број узорака

Варијанса популације

Варијанса популације расподеле података се рачуна готово идентично као варијанса узорака уз разлику да се сума квадратних разлика дели са n – вредношћу броја узорака у целој популацији.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Слика 29: Формула варијансе популације

где је: σ^2 – варијанса популације, x_i – вредност i-тог елемента популације, μ - средња вредност популације, N - број узорака

6. Примери из практичног дела семинарског рада

Практични део семинарског рада обухвата примену описаних техника провера квалитета података над три независна скупа података:

- Скуп 1 – подаци о временским приликама, државним празницима и количини људи у метроу у датом тренутку
- Скуп 2 – подаци са упитника студентима и тинејџерима о личним афинитетима и навикама
- Скуп 3 – подаци о финансијском стању корисника банака који напуштају или не напуштају постојећу банку

Основна идеја практичног дела семинарског рада јесте испитивање квалитета података над описаним скуповима како би се добили детаљни подаци о квалитативним својствима посматраних података у циљу спровођења даљих поступака у фази предобраде података.

Процедуре обављене над свим скуповима ради утврђивања квалитативних својстава посматраних скупова података су:

- израчунавање мера централне тенденције нумеричких и/или категоричких атрибута
- графички приказ расподела података различитих атрибута и уочавање вредности ван опсега
- израчунавање матрице корелације како би се установиле законитости/услољености које делују над подацима
- израчунавање варијансе свих нумеричких атрибута како би се проверила вредност одступања свих вредности од мера централне тенденције
- примена постојеће класе „ProfileReport“ како би се обавила дескриптивна анализа података

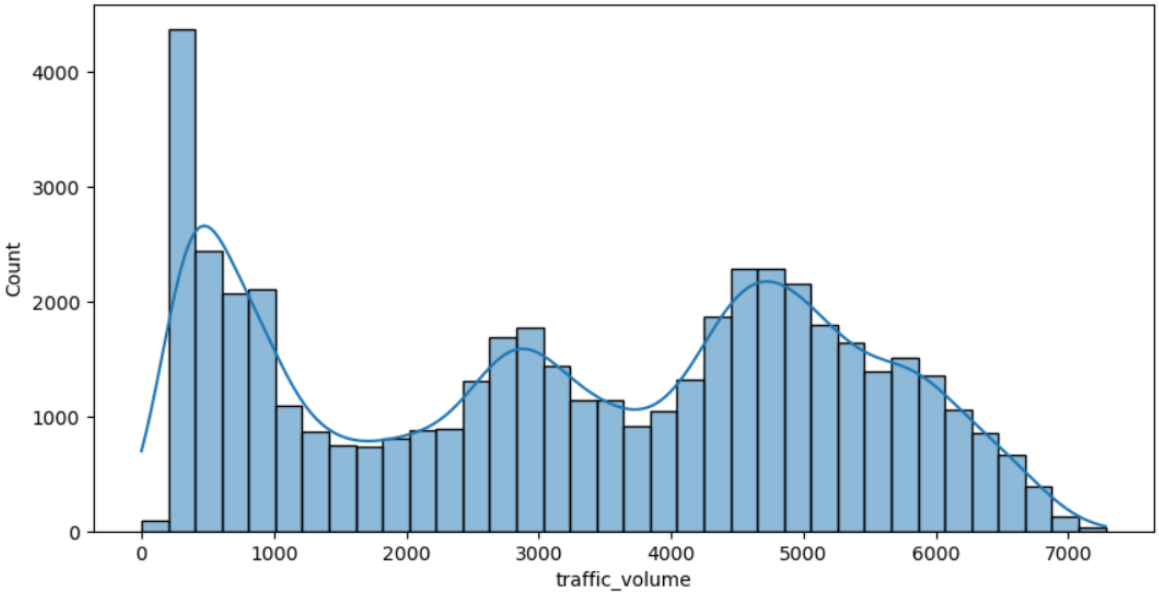
Резултати и закључци над првим скупом података

- *Traffic_volume* атрибут је зависни атрибут овог скупа података са високим опсегом вредности при чему овај атрибут у расподели садржи три „врха“ односно представљен је тримодалним типом расподеле без „outlier“ вредности што се може видети на слици 30.

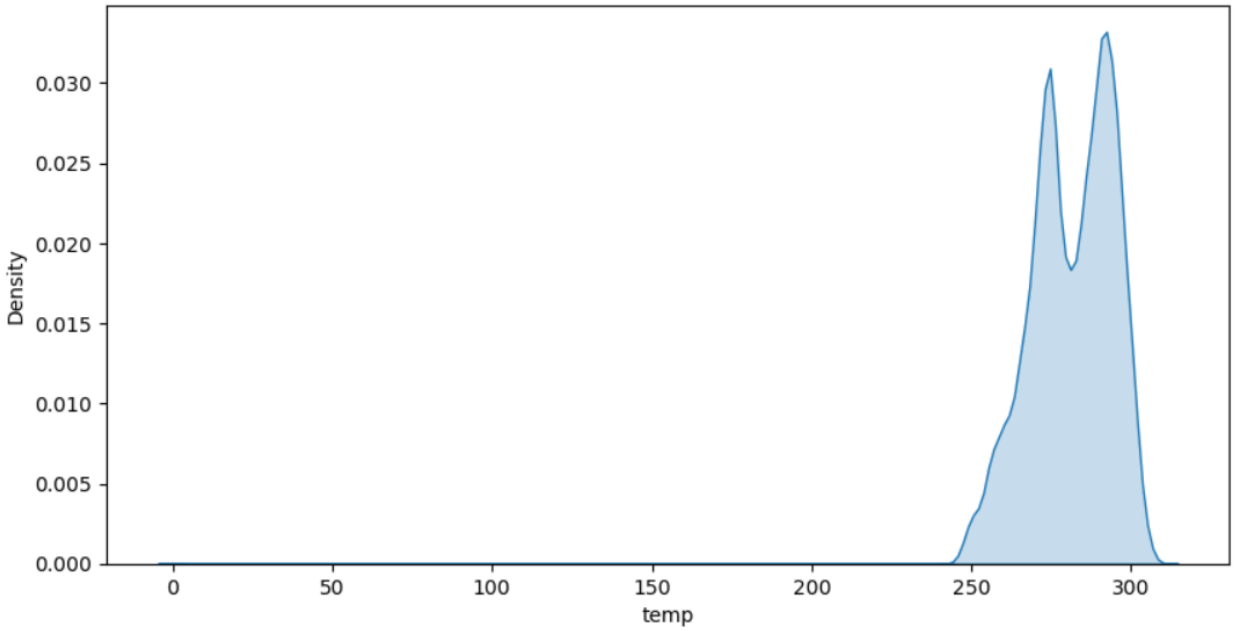
- *Temp* атрибут представља независан атрибут који представља температуру изражену у Келвинима. Овај атрибут је представљен бимодалним типом расподеле и садржи „outlier“ вредности. Ове вредности се уочавају на графику

расподеле података (слика 31) и неопходно је све редове са таквим вредностима избацити из скупа података.

- *Holiday* атрибут представља категорички атрибут са небалансираним вредностима што осликава реалну стварност обзиром да је учесталост празника у току године веома мала у односу на дане када нису празници. Ови небалансирани подаци се могу избацити скупа података или се могу обједити под заједничким ново-изведеним атрибутом који може носити логичку вредност да ли је празник или не.



Слика 30: График расподеле података за атрибут - *Traffic_volume*



Слика 31: График расподеле података за атрибут – *Temp*

- Корелација постоји између парова атрибута: (*traffic_volume, temp*) и (*rain_1h, clouds_all*), али су вредности мале тако да се ове зависности морају пажљиво користити
- Варијанса података је је највећа за атрибуте: *traffic_volume, rain_1h, clouds_all* те се стога мора водити рачуна да не дође до overfitting-а приликом креирања модела машинског учења

Резултати и закључци над другим скупом података

- На основу резултата расподеле података различитих карактеристика испитаника могуће је донети закључке да се ради о групи тинејџера са приближно асиметричним графицима везаним за њихову висину, тежину и број година. Подаци о полу су релативно равномерно распоређени, а највећи број испитаника је некада пробало цигарете. Највећи удео су средњошколски ученици који неколико сати дневно играју игрице.
- Што се тиче корелације између појединих навика и афинитета, може се јасно уочити повезаност између парова атрибута попут (Age, Education, 0.5), (Politics, History, 0.3), (Mathematics, Physics, 0.5), (Shopping, Shopping centres, 0.6) као и (Religion, God, 0.4)
- Обзиром на природу података који се налазе у опсегу вредности 1-5, варијанса података није нарочито употребљива карактеристика на основу које се може донети одређени закључак

Резултати и закључци над трећим скупом података

- Расподела података *CreditScore* атрибута има приближно Гаусову расподелу уз велику учесталост вредности већих од 840
- Код атрибута *Age* примећује се десно кошење асиметричне криве што може изузетно да утиче на развој модела машинског учења
- Атрибут *Tenure* има готово униформну расподелу и не може значајно утицати на исход резултата модела машинског учења, што се види и из матрице корелација јер нема зависности са осталим вредностима
- Атрибут *Balance* има скоро идеалну Гаусову расподелу уз одређен број вредности изван опсега, које НЕ представљају „Outlier" у овом случају
- На матрици корелације се могу уочити парови следећих зависности (*Exited, Age*, 0.27), (*NumOfProducts, Balance*, -0.27), (*Exited, IsActiveMember*, -0.16) и (*Exited, NumOfProducts*, -0.12)
- Вредности варијансе, као и стандардне девијације имају веома високе вредности због своје природе и треба бити изузезетно опрезан приликом фазе тренирања модела како не би дошло до претренираности модела – *overfitting*-а

7. Закључак

У овом раду су теоријски обрађени концепти везани за квалитет података који представљају основну грађу система машинског учења. На основу квалитетних и успешно обрађених података могуће је извести валидне и квалитетне закључке или зависности које делују између података и на тај начин креирати успешан модел машинског учења.

Поступак прикупљања и предобраде података је први поступак у процесу развоја модела машинског учења који има за циљ да изврши основну анализу и преглед улазног скупа података како би се установио квалитативни ниво улазног скупа података.

На основу анализе улазног скупа података као првог поступка обраде података могуће је утврдити скривене информације о својствима података на основу којих се даље може утврдити да ли посматрани скуп репрезентативно осликава реалне вредности које се могу добити у процесу прикупљања података.

Обзиром да квалитет података игра кључну улогу у процесу обучавања и тренирања модела машинског учења, неопходно је најпре извршити одговарајућу анализу и припрему добијеног скупа података како би се установио квалитативни ниво посматраних података.

Одређени предуслови које подаци морају да испоштују укључују: испитивање мера квалитета података, анализа расподеле података, корелације и варијансе. Ови предуслови представљају првобитну информацију о могућностима које се могу касније спровести над тим посматраним скупом података.

Варијанса и корелација представљају статистичке концепте и чине битну улогу у процесу анализе података код машинског учења. Варијанса се односи на меру дисперзије података и примењује се код процене стабилности улазних података над којима се модел обучава. Корелација се користи код извођења међусобних зависности између појединачних фичера у посматраном скупу података и ближе објашњава условљености које делују између података. На основу резултата анализе спроведених помоћу ових статистичких концепата, примењују се различите методе машинског учења како би скуп података адекватно и репрезентативно представљао највећи скуп могућих вредности које се могу појавити.

Разумевање квалитета и расподеле података, варијансе и корелације између појединачних фичера јесте кључни предуслов како би се машинско учење применило на одређени домен проблема који се решава. Стога је неопходно детаљно познавање ових концепата како би се разумео процес креирања модела машинског учења са прецизним предикцијама.

У зависности од домена примене, концепти обрађени у овом раду се примењују на различите начине како би се што боље и успешније обавио процес оптимизације улазног скупа података у циљу креирања прецизнијег модела машинског учења.

8. Литература

1. Salvador García, Julián Luengo and Francisco Herrera, Data Preprocessing in Data Mining, 2015, Springer
2. Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, Third Edition, 2012, Morgan Kaufmann
3. Jason Brownlee, Statistical Methods for Machine Learning, 2018, Machine Learning Mastery
4. Galli, S., Python Feature Engineering Cookbook: Over 70 Recipes for Creating, Engineering, and Transforming Features to Build Machine Learning Models, 2nd Edition, 2022, Packt Publishing, Limited
5. https://www.heavy.ai/technical-glossary/data-quality
6. https://www.kdnuggets.com/2019/10/5-classification-evaluation-metrics-every-data-scientist-must-know.html
7. https://www.metaplane.dev/blog/data-consistency-definition-examples
8. https://dataladder.com/10-data-quality-metrics-you-should-measure/
9. https://www.precisely.com/blog/data-quality/5-characteristics-of-data-quality
10. https://datasciencedojo.com/blog/types-of-statistical-distributions-in-ml/
11. https://towardsdatascience.com/log-normal-distribution-a-simple-explanation-7605864fb67c