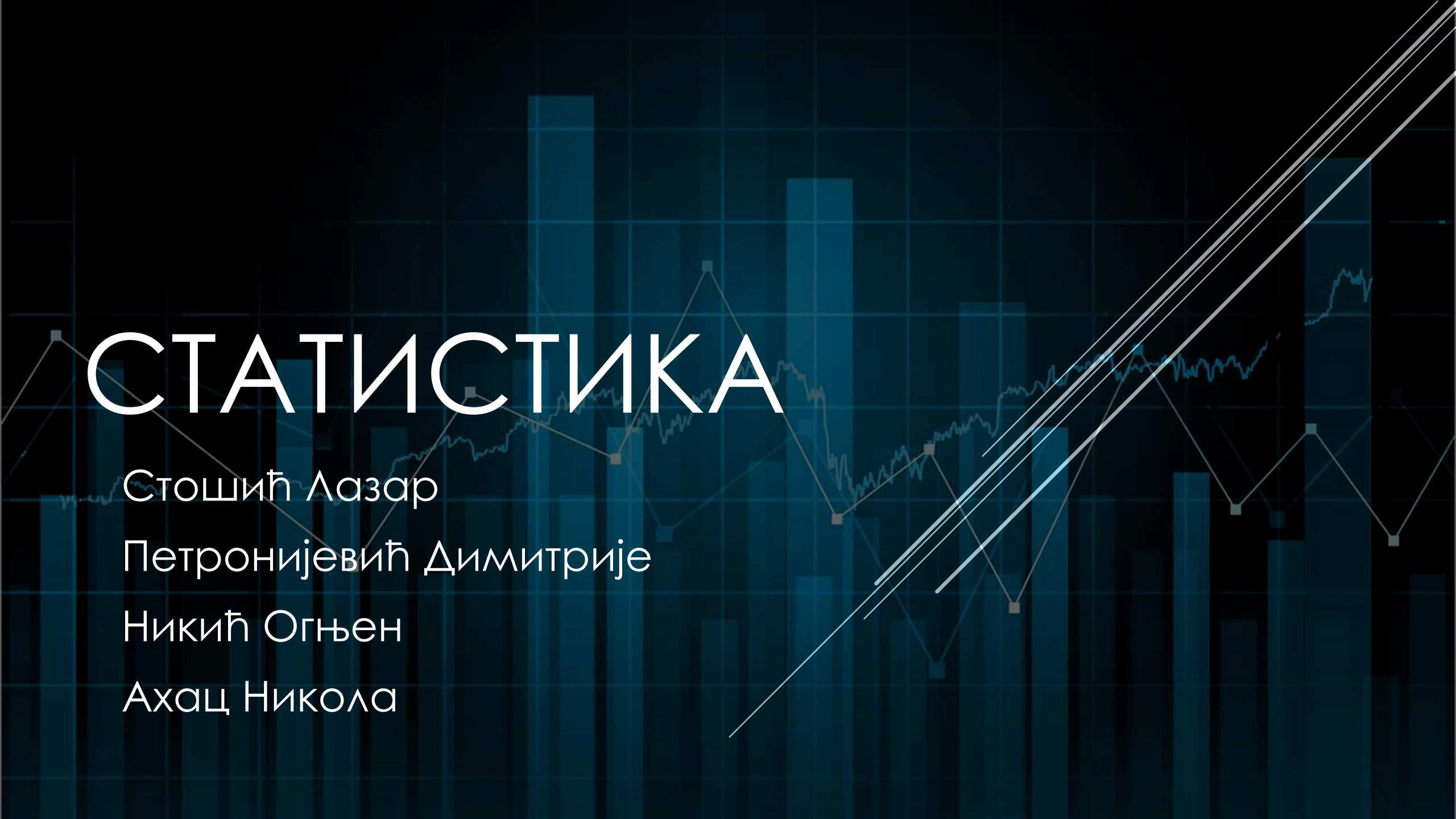




СТАТИСТИКА

Стошић Лазар

Петронијевић Димитрије

Никић Огњен

Ахац Никола

СТАТИСТИКА

- ▶ **Статистика** је област математике која се бави сакупљањем, анализом, интерпретацијом, објашњавањем и презентацијом података.
- ▶ Основна подела је на:
 - ▶ Дескриптивну
 - ▶ Инференцијалну
- ▶ Друга подела би била на:
 - ▶ Параметријску
 - ▶ Непараметријску

Дескриптивна:

- ▶ Бави се мерама централне тенденције (**аритметичка средина, медијана и мод**)
- ▶ Бави се мерама варијабилитета (распон, **стандардна девијација**, варијанца, интерквартилни распон, семиинтерквартилни распон и просечно одступање)
- ▶ Графичко и табеларно приказивање

Инференцијална:

- ▶ Проверавање постављених хипотеза (нултих и афирмативних/алтернативних), уз помоћ статистичких тестова, коефицијената и њихове значајности (т-тест, анализа варијанце, хи-квадрат тест, коефицијенти асоцијације и корелације, дискриминациона анализа, Ман-Витнијев тест, Тест знака...)

Параметријска:

- ▶ Прорачуни се темеље на нормалној (Гаусовој) дистрибуцији

Непараметријска:

- ▶ Прорачуни не морају да се темеље на нормалној (Гаусовој) дистрибуцији

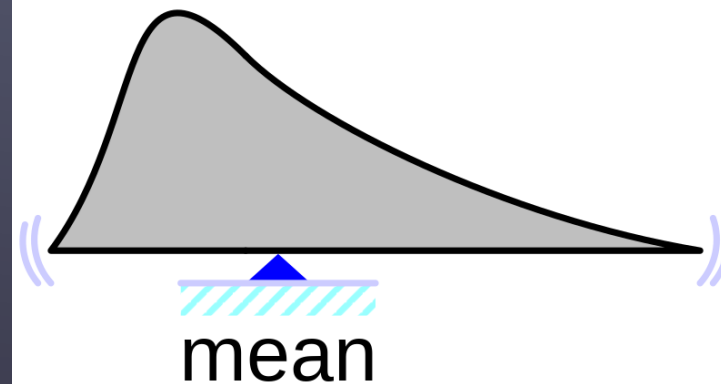
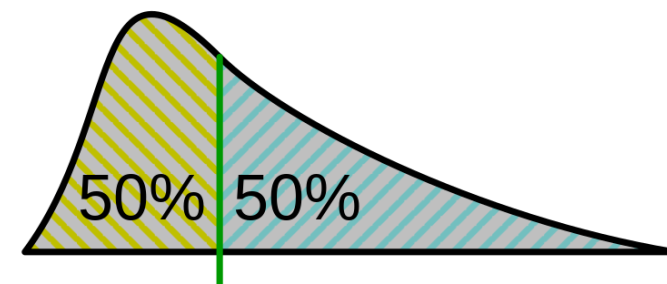
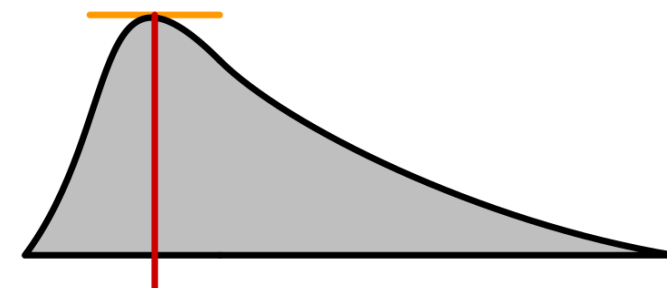
- ▶ Како би се применила нека од статистичких техника/процедура, потребно је прво поставити адекватну хипотезу. Хипотезе могу бити нулте (где се не претпоставља разлика између две или више група испитаника или се не претпоставља да ће корелација између неколико варијабли бити статистички значајна).
- ▶ Примери нултих хипотеза:
 - ▶ *Мачке **немају склоност према одређеној храни** у зависности од њеног облика.*
 - ▶ *На раст биљака **не утиче** боја светлости.*
 - ▶ *Музичке способности **нису директно везане** за старост.*
- ▶ Такође, постоје и афирмативне хипотезе, којима се претпоставља нека статистички значајна разлика или повезаност.
- ▶ Примери афирмативних хипотеза:
 - ▶ *Количина подебљаног теста у књигама/скриптама **је повезана** са резултатима на испитима.*
 - ▶ *Коришћење конца за зубе **има ефекта** на количину каријеса*

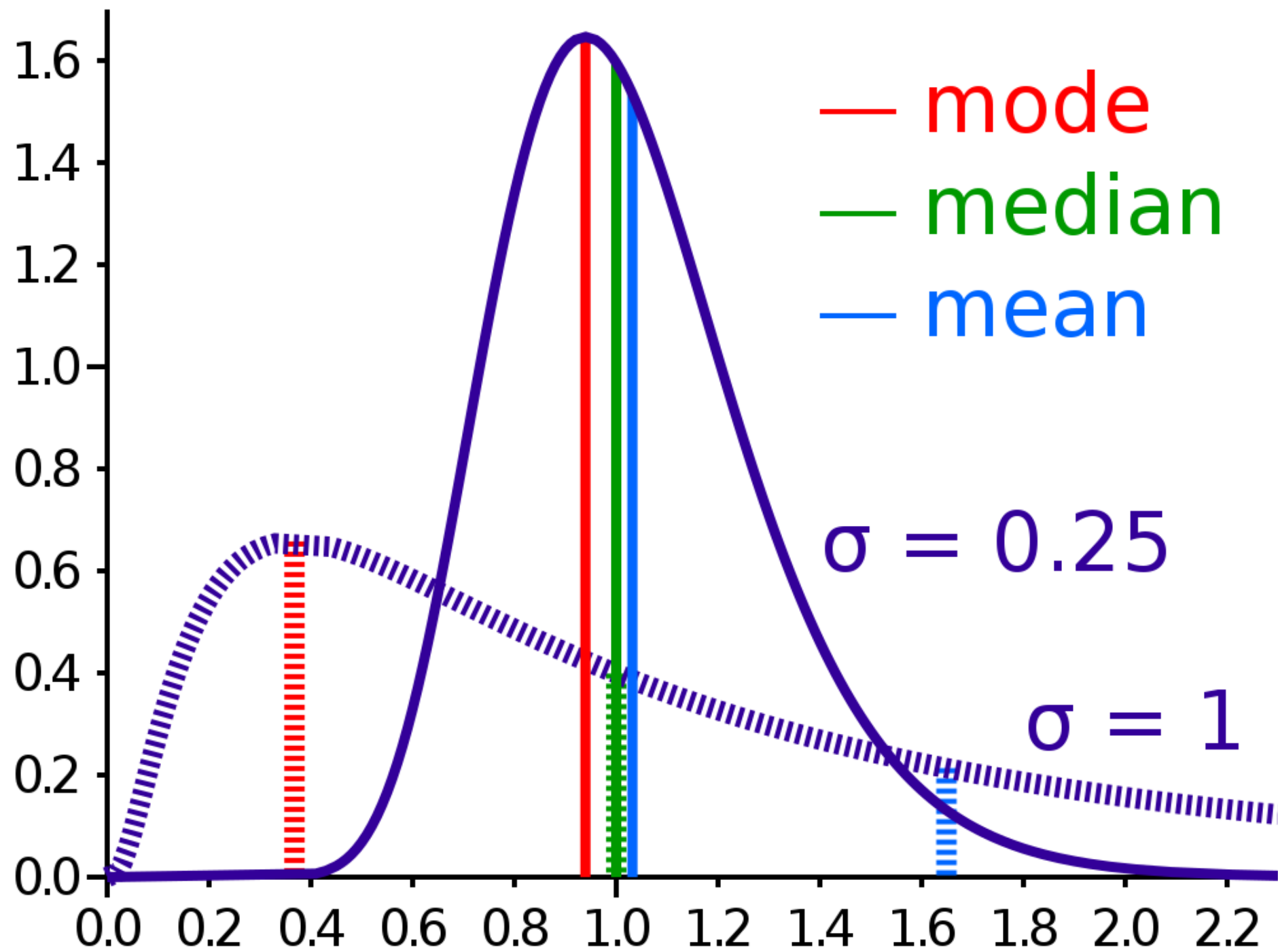
АРИТМЕТИЧКА СРЕДИНА, МЕДИЈАНА И МОД

- ▶ **Аритметичку средину** или **просек** неког скупа бројева дефинишемо као количник суме свих вредности тог скупа и броја његових елемената.
- ▶ Као и друге мере централне тенденције, представља неку вредност око које се остале вредности "окупљају".
- ▶ За разлику од осталих мера тенденције, аритметичка средина доста зависи од екстремних вредности

- ▶ **Медијана** неког скупа бројева се дефинише као вредност која горњу половину вредности од доње половине, тј. средњи члан низа који се добија када сортирамо све елементе скупа.
- ▶ Ако низ има непаран број елемената, узимамо средишњи елемент као медијану.
- ▶ Ако низ има паран број вредности, онда је медијана ар. средина од средишња два елемента.
- ▶ Није осетљива на екстремне вредности, тако да у скупу где постоји више екстремних вредности, може понекада боље приказати које су "типичне" вредности за тај скуп

- ▶ **Мод** неког скупа је просто вредност која се најчешће у њему појављује
- ▶ Ретко се користи, и најчешће за номиналне вредности, тј. вредности које нису бројеви, већ имају одређена имена
- ▶ За разлику од просека и медијане, у зависности од скупа може постојати више модова, а не мора постојати ни један.



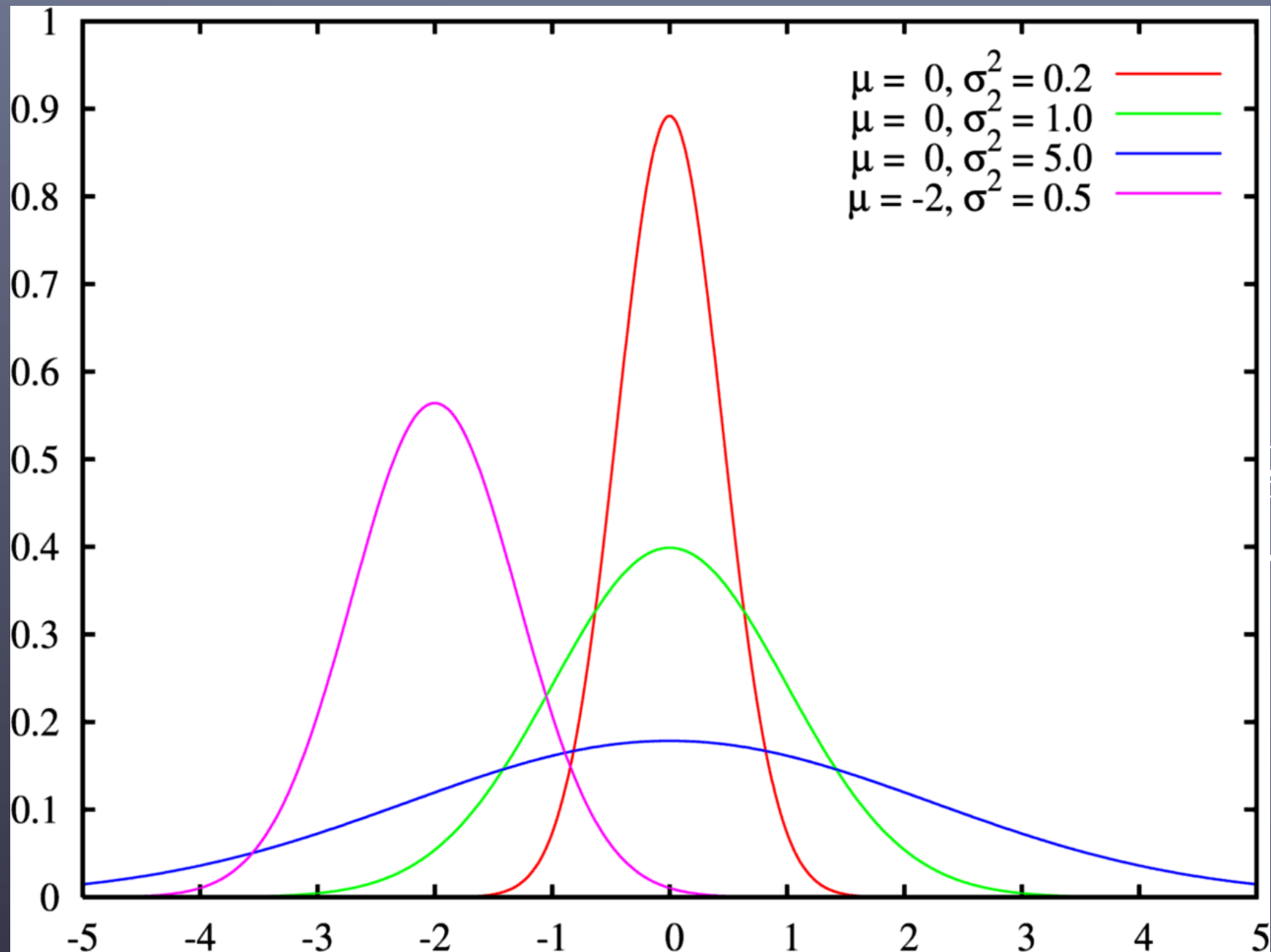


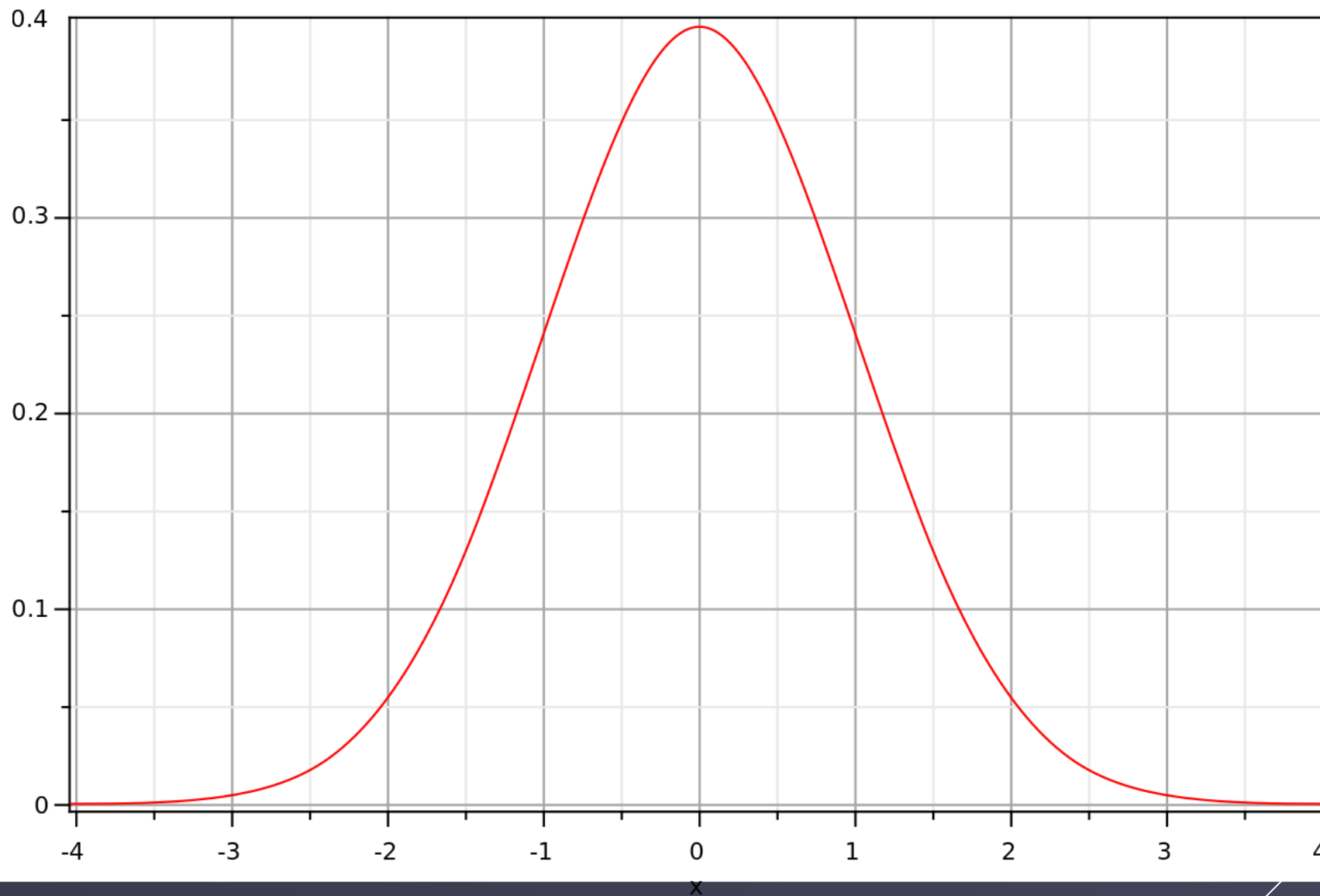
НОРМАЛНА РАСПОДЕЛА

- ▶ **Нормална расподела**, некада звана и **Гаусова расподела**, спада у непрекидне расподеле вероватноће.
- ▶ Чланови фамилије нормалне расподеле су дефинисани преко два параметра, математичко очекивање, и варијанса (дисперзија) σ^2 .
- ▶ Главна одлика ових расподела је да означи колики број података тежи одређеној вредности.
- ▶ Нормална нормирана расподела је нормална расподела са очекивањем једнаким нули, и варијансом једнаком један (зелена функција на следећем слајду).
- ▶ Значај ове расподеле огледа у многим мерењима и физичим феноменима, јер њихови резултати већински могу апроксимирати овом расподелом.

Општа формула
функција непрекидне
расподеле вероватноће:

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



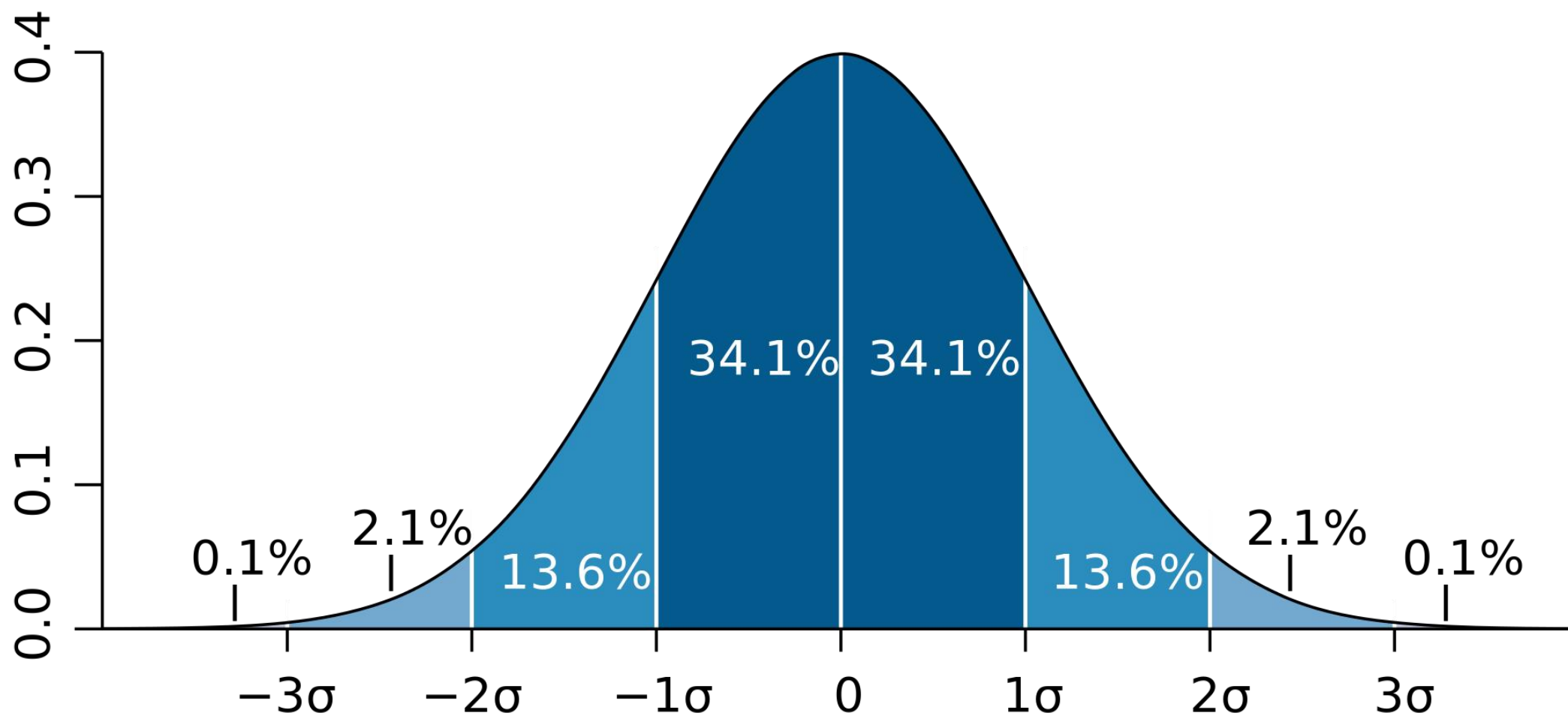


Горе: график (Гаусове) нормалне расподеле
Десно: формула функције по којој се рачуна
нормална расподела

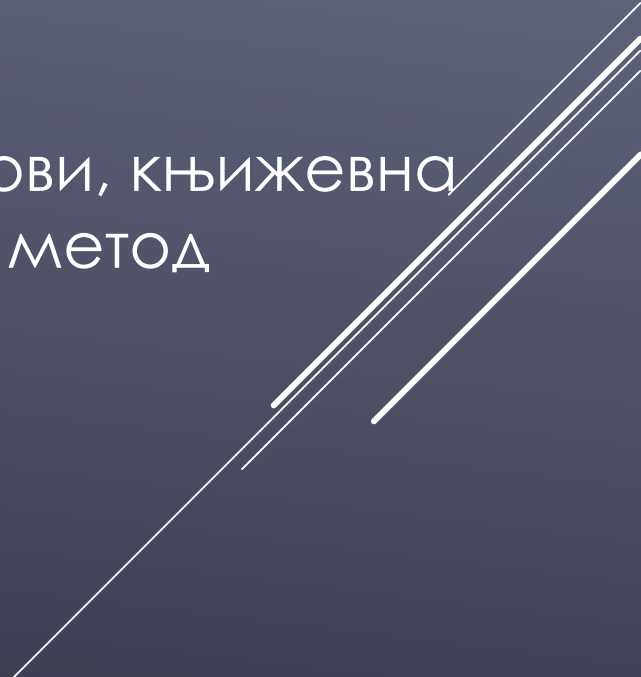
$$\varphi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

СТАНДАРДНА ДЕВИЈАЦИЈА

- ▶ Стандардна девијација је у статистици апсолутна мера дисперзије у основном скупу. Она говори, колико у просеку елементи скупа одступају од аритметичке средине скупа. Означава се грчким словом сигма, σ .
- ▶ Ниска стандардна девијација указује на то да вредности имају тенденцију да буду близу средње вредности (која се назива и очекиваном вредношћу) скупа, док висока стандардна девијација указује да су вредности распоређене у ширем опсегу.
- ▶ Ако неки подаци одговарају нормалној расподели података, онда на удаљености од 1σ од аритметичке средине би требало да се налази приближно 68% података. На 2σ се налази 95,4%, на 3σ 99,7%... Ово називамо **правилем 68-95-99,7** или **емпиријским правилем**.
- ▶ За нормалну расподелу, две тачке на кривој које су удаљене једну стандардну девијацију од криве су такође и превојне тачке.



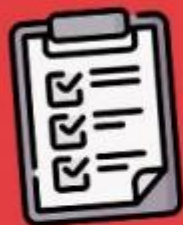
АНКЕТЕ И УЗОРЦИ

- ▶ Анкета или упитник је техника прикупљања података и обавештења која се примењује у друштвеним наукама.
 - ▶ Већи број испитаника неке анкете чине узорак.
 - ▶ Термином анкета су обухваћени новинарски жанрови, књижевна анкета и анкета уопште, којом се у науци означава метод прикупљања чињеница.
- 
- Three decorative white lines of varying lengths and slopes are positioned in the bottom right corner of the slide.

- ▶ Бирање узоракa је веома важан и тежак део статистике и самог анетирања. Да би нека наша претпоставка била тачна, наш узорак мора да буде претставник групе коју ми анкетирамо.
- ▶ За анкету је обавезно да има објективан и довољно велики узорак, како би цело истраживање било валидно.
- ▶ Нпр.
 - ▶ Ако при оцењивању супе пробамо само кап, наш узорак ће се сматрати превише мали.
 - ▶ Са друге стране, ако напунимо целу кашику, али захватимо само део супе у коме је кувар ставио доста бибера, наш узорак ће бити необјективан и наша претпоставка нетачна.
- ▶ Узорак не треба бити ни превелики, зато што се тај вишак прецизности који доноси већи узорак не исплати.

- ▶ Поред величине, одабир узрока је још једна веома важна ставка анкетирања.
- ▶ Најефикаснији одабир је тај којим се бира потпуно насумично, тиме да свако има једнаку шансу да буде изабран, и да било који одабир не утиче на даље бирање узорка.
- ▶ Иако техника да само насумично бирамо појединце није савршена, из ове идеје су уследиле разне савремене технике одабира, нпр. бирање раслојавањем, кластер бирање и квота бирање, док се у сваком поштује принцип насумичног бирања до неке мере.

Циљана група



Узорак



ПРИМЕРИ АНКЕТА

- ▶ Једна радио станица је поренула анкету о томе да се легализује марихуана. Јавило се 3636 слушалаца, и од њих, 78% се слаже са том идејом. Станица после резултата тврди да је време да се марихуана легализује и након тога контактира владу. У овом случају, видимо да узорак није валидан јер није насумичан, чинећи само слушаоце тог радија, још поред тога само они људи који су одлучили да се јаве. Тиме се није могло закључити мишљење целе популације на ову тему.

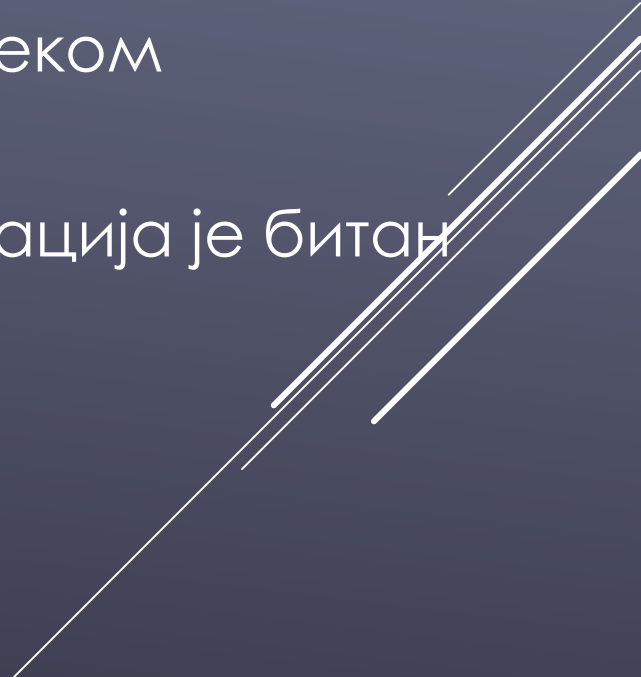
- Пре неколико година, једна анкета компаније Gallup на Америчким универзитетима је утврдило да 33% америчких студената не зна метрички систем. Са друге стране, новинарски чланак у дневним новинама Калифорније је такође анкетирао читаоце и тиме закључио да 98% американаца заиста зна метрични систем. Анкета у новинама је очигледно субјективна, јер је врло вероватно да су се јавили само људи који су заиста знали метрички систем.



- 2 000 људи је било одабрано да одговори на одређени број питања. Анкета се вршила преко телефона и изабрани су насумични пројеви теефона из базе података свих бројева телефона. Ово је на први поглед одличан одабил узорака, али сетимо се да постоји сиромашан слој друштва који није у могућности да пријушти мобилни телефон, и тиме ова анкета постаје невалидна.

- ▶ Следећи битан фактор анкетирања је одабир питања анкете.
- ▶ Питања неке анкете веома лако могу да постану двосмислена у зависности од тога коју групу људи питате, формулацију реченице, колико су питања прецизна итд.
- ▶ Питања анкете је веома тешко саставити јер, као што је речено, веома лако може доћи до заблуде при одговарању.
- ▶ Нпр. Ако поставимо питање „Да ли читате новине?“, можда звучи као прецизно и разумно питање, али оно се може протумачити на неколико начина. Да ли читате новине често? Понекад? Сваки дан? Да ли прочитате све? итд.
- ▶ Такође, одговор на питање „Да ли конзумирате много алкохола“, варира од тога шта наш субјекат сматра под „алкохол“ и „много“, чиме се драстично мењају резултати истраживања.

ШТА ЈЕ КОРЕЛАЦИЈА?

- ▶ Корелација је повезаност вредности две варијабле
 - ▶ Ако знамо вредност једне варијабле, можемо са неком сигурношћу да предвидимо вредност друге
 - ▶ Налажење, проверавање и квантификовање корелација је битан део статистике
- 
- Several thin, parallel white lines of varying lengths and angles are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

- ▶ Корелације су корисне јер могу да нам указују на предвидљив однос неких варијабли, који се у пракси може искористити
- ▶ На пример, постоји корелација између потражње за струјом и временских прилика
- ▶ У овом случају то није само корелација него је и узрочност, јер екстремне временске прилике доводе до тога да људи морају да користе више струје да би се грејали или хладили

- ▶ Ако између две варијабле постоји корелација, не значи да између њих постоји и узрочност
- ▶ Постојање корелације између А и В може да значи различите ствари:
 - ▶ да а проузрокује б
 - ▶ да б проузрокује а
 - ▶ да су а и б потпуно случајно повезани
 - ▶ да трећи фактор Ц утиче истовремено и на а и на б, а да а и б сами по себи немају никакву везу

На пример:

- ▶ Замислимо студију која показује да је конзумирање марихуане у средњим сколама и факултетима повезано са нижим оценама
- ▶ То се може објаснити на више начина:
 - ▶ Коришћење марихуане проузрокује ниже оцене
 - ▶ Имање нижих оцена наводи студенте на то да мање брину о свом академском успеху и да се горе осећају, па то доводи до коришћења марихуане
 - ▶ Социјалнији студенти истовремено имају веће шансе да конзумирају марихуану и да мање важности придодају оценама

ПРИМЕРИ ОДНОСА КОРЕЛАЦИЈА И УЗРОЧНОСТИ

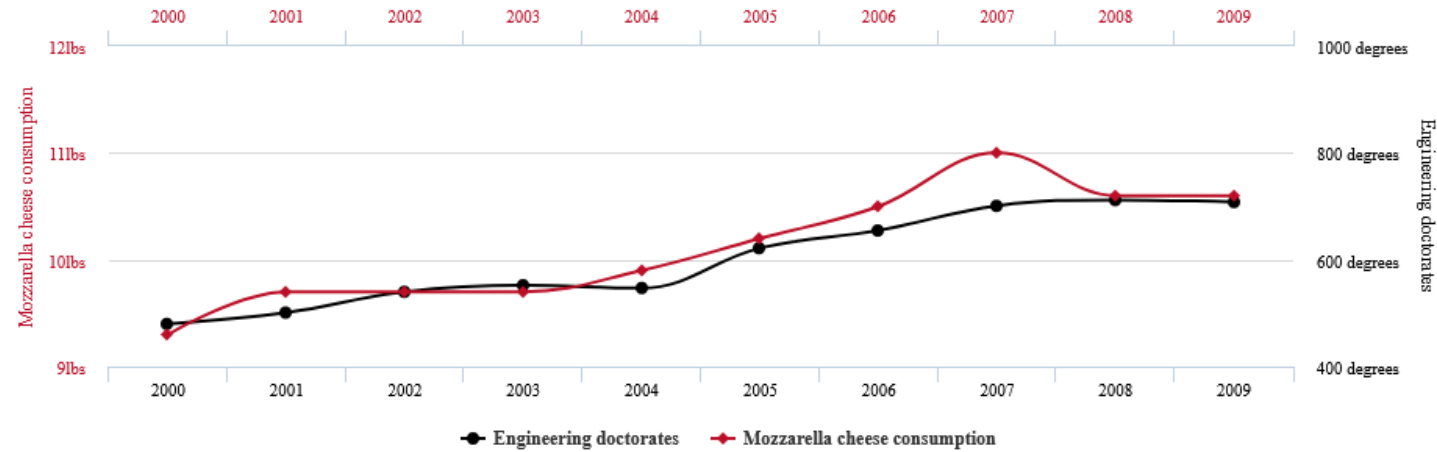
- ▶ У неким државама је примећено да број рода на крововима кућа има јаку корелацију са бројем деце који живе у тим кућама
- ▶ Иако је корелација присутна, не мора да значи да роде проузрокују да се рађа више деце
- ▶ Боље објасњење је то што куће у којима живе веће породице обично имају веће кровове, па зато привлаче и већи број рода
- ▶ Често се између потпуно неповезаних података може наћи корелација ако они имају сличне вредности

Per capita consumption of mozzarella cheese

correlates with

Civil engineering doctorates awarded

Correlation: 95.86% ($r=0.958648$)

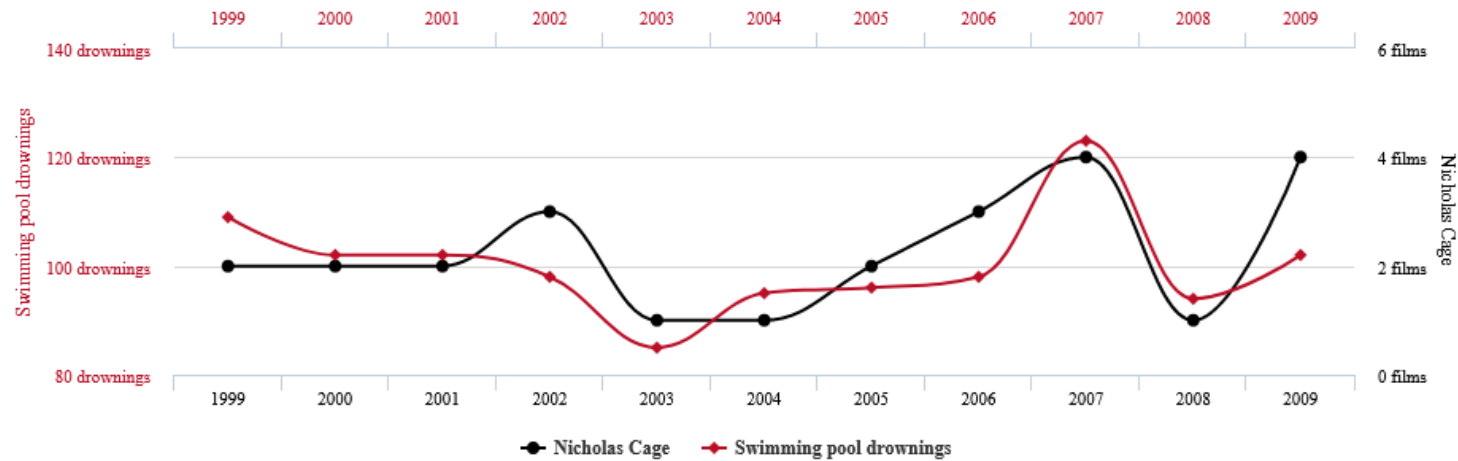


Number of people who drowned by falling into a pool

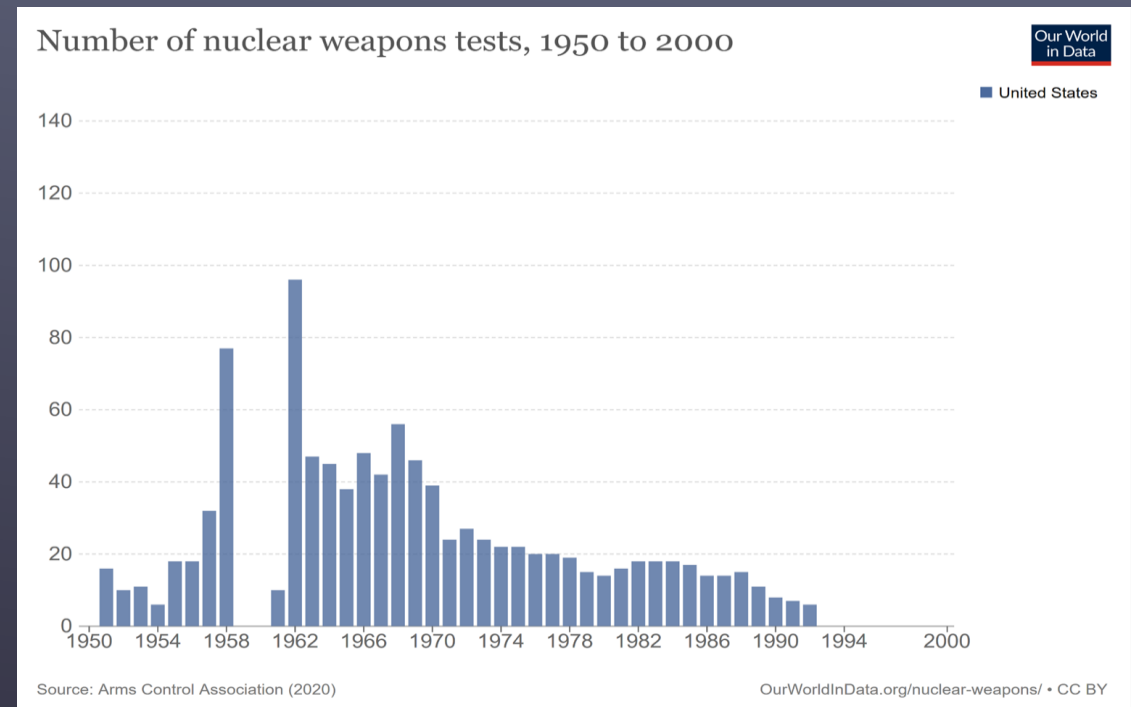
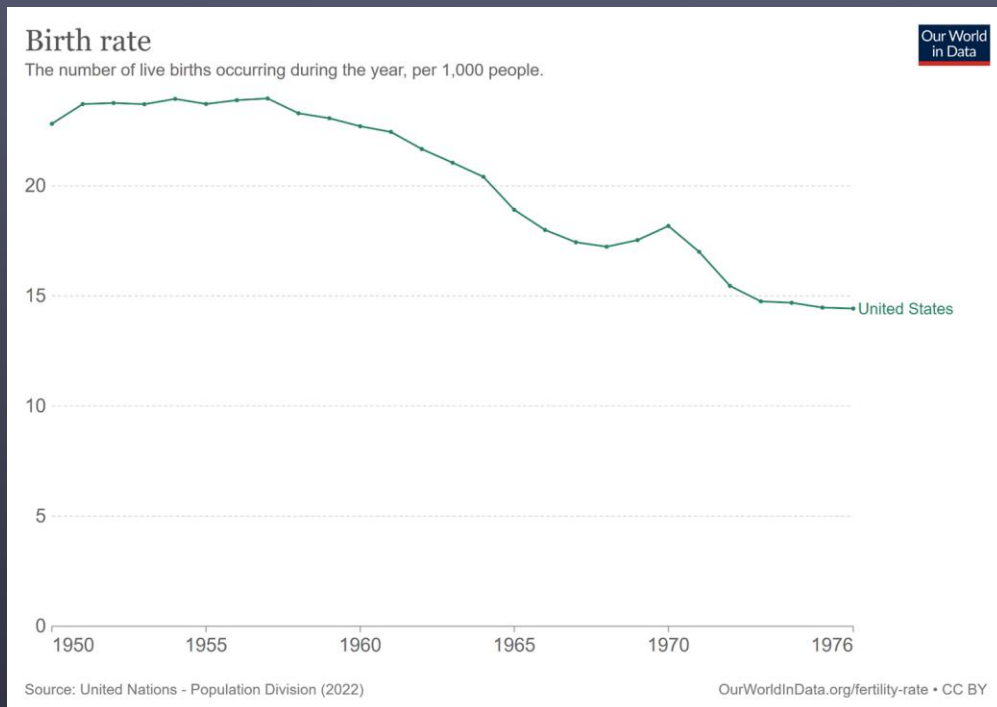
correlates with

Films Nicolas Cage appeared in

Correlation: 66.6% ($r=0.666004$)



- ▶ Некада може да се деси да између неких података не постоји корелација, чак иако нам делује да постоји
- ▶ Сјајан пример је то што је око 1960-их почео да пада наталитет у америци, што је такође време када су тестирања нуклеарних оружија била на врхунцу



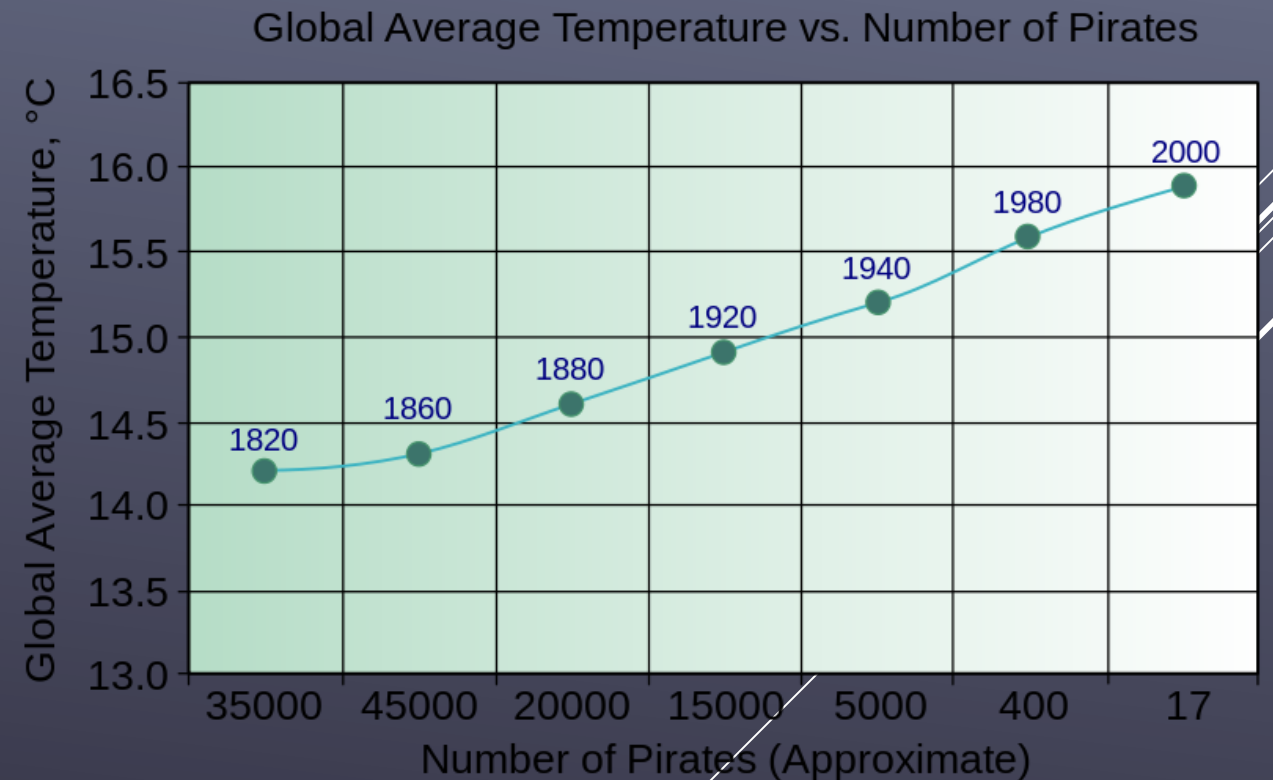
- ▶ То можда би можда деловало као узрочност, али је више вероватно објашњење то што су 1960-их у америци у продају кренуле да се уводе контрацептивне пилуле
- ▶ Упркос овоме, идаље је могуће да су многи други фактори одиграли улогу у опадању наталитета

МЕШАЊЕ КОРЕЛАЦИЈА И УЗРОЧНОСТИ

- ▶ Мешање ова два појма може да доведе до лоших закључака, као и до стварања сујеверја
 - ▶ На пример, монаси у кини су веровали да је помрачење сунца уствари један велики пас који покушава да поједе сунце
 - ▶ Да би отерали пса вршили су ритуал који је укључивао ударање у гонг
 - ▶ То што би сваки пут кад ударе о гонг помрачење сунца после неког времена прошло би им потврдило то у шта верују
- ▶ Још један феномен који настаје због ове конфузије је регресија ка средњој вредности

КАКО ИЗБЕЋИ ПОГРЕШНЕ ЗАКЉУЧКЕ

- ▶ Постојање корелације не значи да постоји и узрочност
- ▶ Да би постојала узрочност мора се доказати да вредност неке варијабле стварно утиче на вредност друге
- ▶ Потребно је обратити пажњу на графове којима су презентовани податци



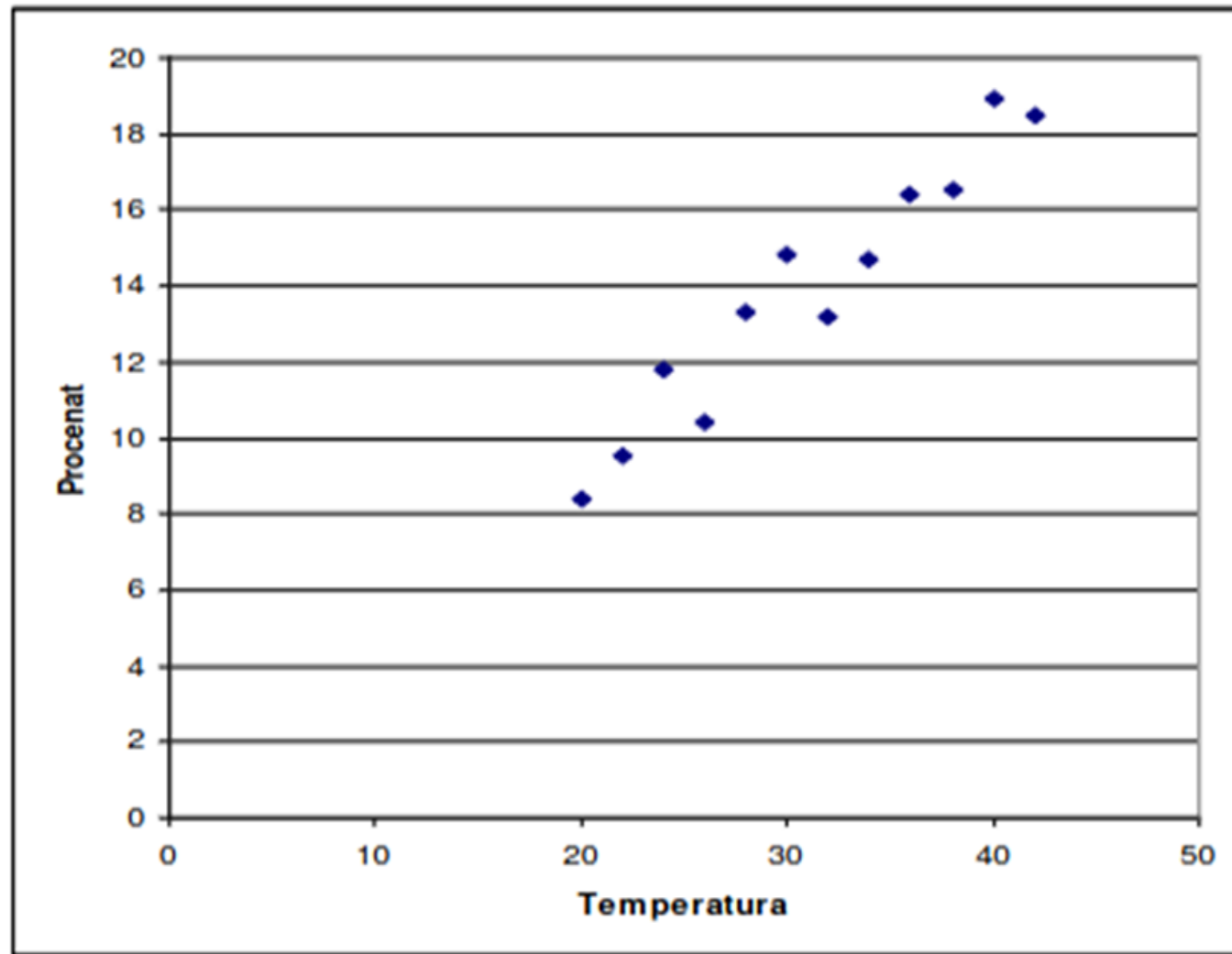
РЕГРЕСИЈА

- ▶ Један од циљева у великом броју истраживања је да се опишу везе међу појавама које нас окружују.
- ▶ То се може постићи проналажењем формуле или једначине која повезује величине које посматрамо.
- ▶ На пример:
 - ▶ веза између температуре и притиска у хемијском процесу
 - ▶ или веза између броја јабука на дрвећу у воћњаку и количине ђубрива које је коришћено у воћњаку

- ▶ У статистици проналажењем статистичких веза између појава бави се регресиона анализа, регресија.
- ▶ Регресија је од великог значаја, како у економији и привреди, тако и у другим природним наукама, као што су: хемија, физика, биологија, фармакологија, токсикологија, биохемија и судска медицина....
- ▶ Пример: У табели 1. дати су подаци о проценту издвојене супстанце за различите висине температуре при неком хемијском процесу.

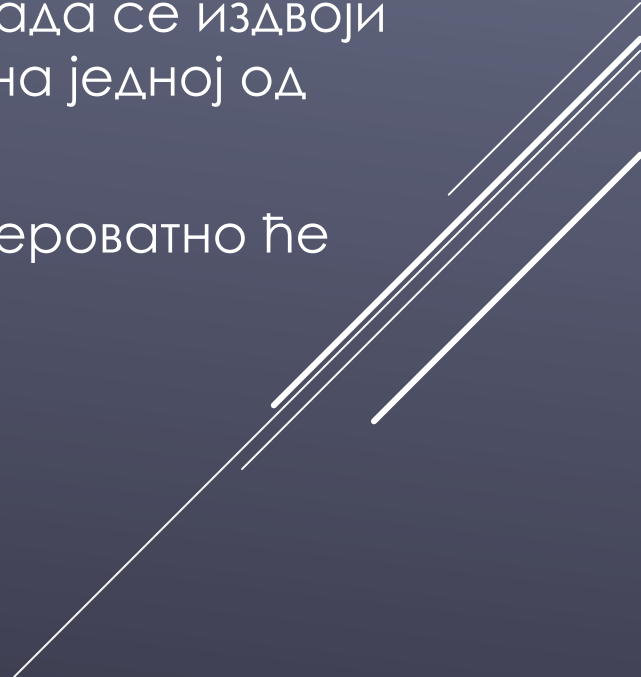
Tabela 1.

temperatura	20	22	24	26	28	30	32	34	36	38	40	42
procenti	8.4	9.5	11.8	10.4	13.3	14.8	13.2	14.7	16.4	16.5	18.9	18.5



Slika 1.

РЕГРЕСИЈА КА СРЕДИНИ

- ▶ "Регресија ка средини" је статистички феномен који се јавља када постоји позитивна или негативна повезаност између две варијабле и када се издвоји група људи са екстремно високим или ниским вредностима на једној од варијабли.
 - ▶ Када се ти исти људи поново испитују након неког времена, вероватно ће имати мање екстремне вредности.
- 
- Several white lines of varying lengths and angles are drawn in the bottom right corner of the slide, creating a modern, abstract graphic element.

ПРИМЕР

- ▶ Појам је увео енглески научник Френсис Галтон (статистичар, социолог, психолог, антрополог...).
- ▶ Испитивао је односе међу висинама очева и синова
- ▶ Открио је следеће:
 1. Високи очеви имају високе синове.
 2. Веома високи очеви имају синове који су нижи од њих (Несавршена колерација 2 променљиве)
 3. Веома ниски очеви имају синове који су виши од њих (Несаврсена колерација 2 променљиве)

SALLY CLARK

- Sally Clark је била британска адвокатица која је била неправедно оптужена И осуђена за убиство своје две бебе у раним 90-им годинама прошлог века.



FBI АНАЛИЗА КОСЕ

- ▶ Један од примера погрешне примене регресионе анализе у правосуђу је био случај " FBI анализа косе" у Сједињеним Државама.
- ▶ Овај случај указује на важност правилне примене статистичких метода у правосуђу и на опасности од претераног ослањања на једну врсту доказа. У том смислу, важно је да се у правосудним случајевима користе различите методе и докази како би се избегле грешке у закључивању.