

Зететика - Статистичке замке, Случајан догађај

Данијела Симић
зететика - увод
31. март 2021.



1. Увод
2. Статистика
3. Истраживања и узимање узорка
4. Корелација
5. Регресија ка средини и сујеверје
6. Графици и илустрације: вреде као хиљаду лажи
7. Проценти и подвале

Увод

Мало људи је разуме, али сви је уважавају.

Рекли су ... (о статистици)

Бенџамин Дизраели (1804 – 1881), британски државник

Постоје три врсте лажи: обичне лажи, важне лажи, и статистика.

Рекли су ... (о статистици)

Бенџамин Дизраели (1804 – 1881), британски државник

Постоје три врсте лажи: обичне лажи, важне лажи, и статистика.

Херберт Џорџ Велс (1866 – 1946), енглески научнофантастични књижевник

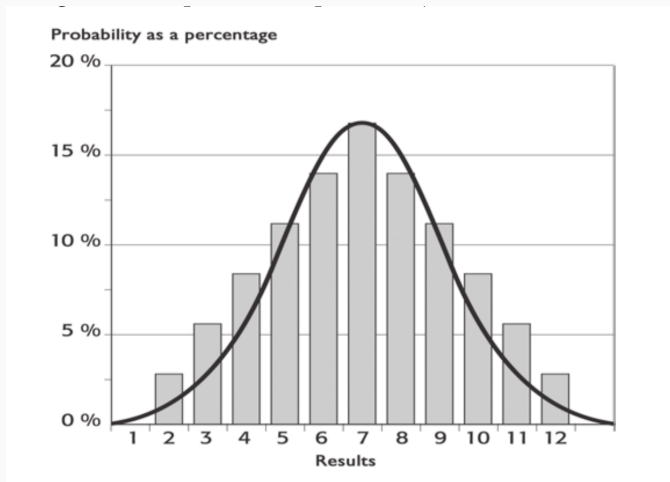
Као што је сада потребно да знамо да пишемо и читамо, тако ће у 20. веку бити потребно да знамо статистику.

Статистика

Статистика је функција узорка (x_1, x_2, \dots, x_n) чији аналитички израз не зависи од непознатих параметара расподеле обележја популације из које је узорак узет.

Гаусова крива – пример

Збир који добијамо бацањем две коцкице.



Нормална расподела или Гаусова расподела.

Најчешће мере које се користе у дескриптивној статистици:

- Аритметичка средина

Најчешће мере које се користе у дескриптивној статистици:

- Аритметичка средина
- Медијана

Најчешће мере које се користе у дескриптивној статистици:

- Аритметичка средина
- Медијана
- Модус

Најчешће мере које се користе у дескриптивној статистици:

- Аритметичка средина
- Медијана
- Модус
- Стандардна девијација

$$\mathit{arth}(X) = \bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\text{arth}(X) = \bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Аритметичка средина се најчешће користи. Узимају се сви подаци у разматрање. Ипак, осетљива је на **екстремне вредности**.

Медијана се у теорији вероватноће и статистици описује као број који раздваја горњу половину узорка, популације или расподеле вероватноће од доње половине.

Медијана се у теорији вероватноће и статистици описује као број који раздваја горњу половину узорка, популације или расподеле вероватноће од доње половине.

Медијана се такође често користи. Не узима у обзор све вредности и није осетљива на екстремне вредности. Понекад може боље да опише податке од аритметичке средине.

Модус је вредност која се у узорку или групи података појављује најчешће.

Модус је вредност која се у узорку или групи података појављује најчешће.

Модус се најређе користи. У неком узорку може бити више модуса или ни један. Не посматра све вредности. Обично се користи код номиналних вредности (**променљиве описане именом**) или код дискретних променљивих (**могу имати само ограничен број вредности**).

Пример

$$X = 109, 129, 129, 135, 139, 149, 159, 179$$

$$\bar{X} = \frac{1128}{8} = 141$$

$$\text{медијана} = \frac{135+139}{2} = 137$$

$$\text{модус} = 129$$

Лажно представљање података

Марко се запошљава у једној компанији. Компанија има шефа, његовог брата и шест рођака. Особље чини десет радника и пет супервизора.

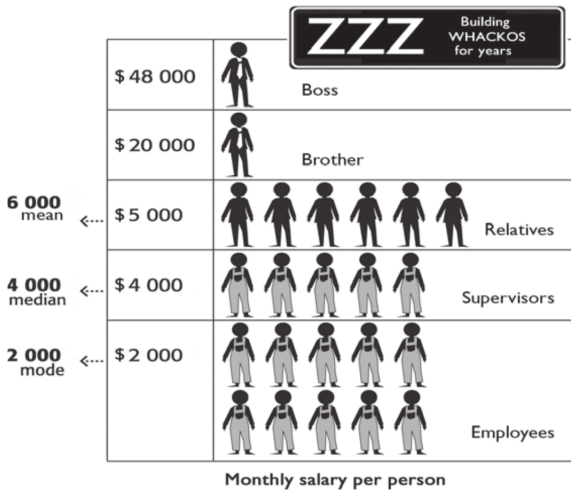
Шеф каже да је просечна плата у фирми 6000 долара месечно. Каже да ће Марко у почетку примати 1500 долара, а да ће после пробног периода плата значајно порастити.

Лажно представљање података

Марко се запошљава у једној компанији. Компанија има шефа, његовог брата и шест рођака. Особље чини десет радника и пет супервизора.

Шеф каже да је просечна плата у фирми 6000 долара месечно. Каже да ће Марко у почетку примати 1500 долара, а да ће после пробног периода плата значајно порастити.

После десетак дана: Марко одлази код шефа - лагао си ме!



You
tricked me
about the
average salary!

You are
much more
intelligent than average,
a big mistake in my
business. You're fired.



На основу аритметичке средине не може се увек извући податак о расподели вредности!

Стандардна девијација

Стандардна девијација је у статистици апсолутна мера дисперзије у основном скупу. Она нам говори, колико у просеку елементи скупа одступају од аритметичке средине скупа.

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

Пример

Отишли сте да пецате на језеро, али Вам је речено да се раније у језеру изливали токсини из оближње фабрике. Рекли су Вам да је за човека штетно ако у риби коју поједе има 7mg токсина. Такође су Вам рекли да у просеку у риби има 4mg токсина.

Пример

Отишли сте да пецате на језеро, али Вам је речено да се раније у језеру изливали токсини из оближње фабрике. Рекли су Вам да је за човека штетно ако у риби коју поједе има 7mg токсина. Такође су Вам рекли да у просеку у риби има 4mg токсина.

- Да ли бисте појели рибу?

Пример

Отишли сте да пецате на језеро, али Вам је речено да се раније у језеру изливали токсини из оближње фабрике. Рекли су Вам да је за човека штетно ако у риби коју поједе има 7mg токсина. Такође су Вам рекли да у просеку у риби има 4mg токсина.

- Да ли бисте појели рибу?
- Да ли бисте појели рибу ако знате да је стандардна девијација 1mg ?

Пример

Отишли сте да пецате на језеро, али Вам је речено да се раније у језеру изливали токсини из оближње фабрике. Рекли су Вам да је за човека штетно ако у риби коју поједе има 7mg токсина. Такође су Вам рекли да у просеку у риби има 4mg токсина.

- Да ли бисте појели рибу?
- Да ли бисте појели рибу ако знате да је стандардна девијација 1mg ?
- А у случају да је стандардна девијација 4mg ?

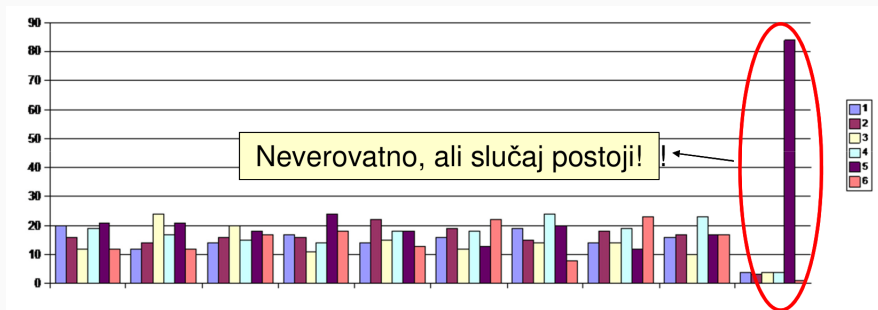
Степен значајности једног статистичког резултата

Вероватноћа да размак (девијација) између:

- резултата посматрања
 - теоријског предвиђања
- буде последица случаја.

Пример: тестирање да ли је коцкица добро балансирана.

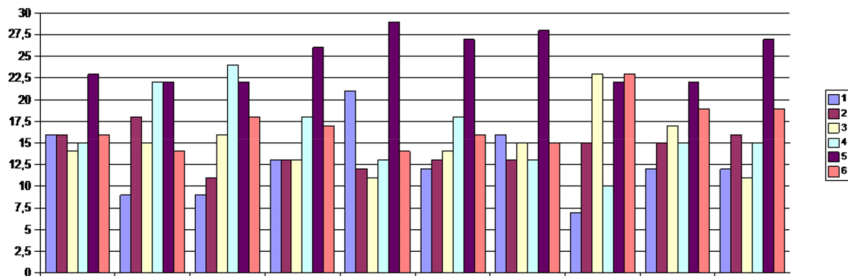
Посматрајмо дистрибуцију (10 серија од по 100 бацања):



Ако је коцка намештена, 5 ће се појавити више пута у свакој серији бацања.

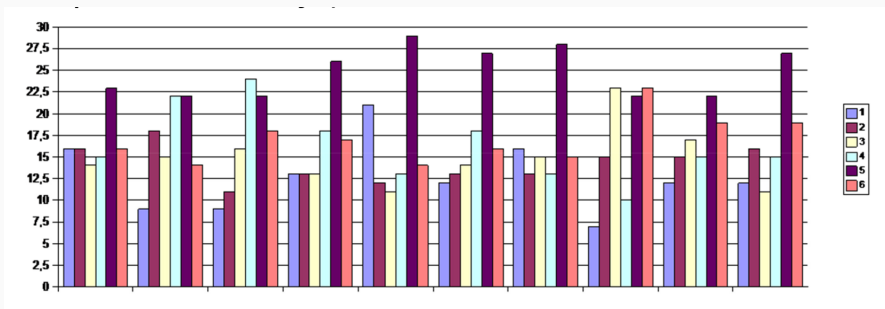
Оглед

Шта закључујемо о следећем огледу (10 серија од по 100 бацања):



Оглед

Шта закључујемо о следећем огледу (10 серија од по 100 бацања):



Овог пута је требало имати среће 10 пута што није разумна претпоставка.

Истраживања и узимање узорка

Статистика нам омогућава да закључујемо о особинама неке популације на основу малог дела те популације који називамо **узорак**.

Истраживања и узимање узорка

Статистика нам омогућава да закључујемо о особинама неке популације на основу малог дела те популације који називамо **узорак**.

Углавном није могуће испитати целу популацију (због времена или новца који један за то потребан). Зато узимамо узорак.

Истраживања и узимање узорка

Статистика нам омогућава да закључујемо о особинама неке популације на основу малог дела те популације који називамо **узорак**.

Углавном није могуће испитати целу популацију (због времена или новца који један за то потребан). Зато узимамо узорак.

Креирање узорака и прављење закључака на основу тих узорака је најраспрострањенија и најважнија **примена статистике**.

Да би закључак о некој популацији био валидан, онда узети узорак мора да буде **представник** те популације.
Да би задовољио овај услов узорак мора да буде **велики** и **непристрасан**.

Пример

Literary Digest је од 1920. спроводио истраживање о победнику Америчких избора. Иако је неколико година био успешан, 1936. резултат истраживања је био погрешан.

Часопис је својим претплатницима слао анонимне упитнике и послао би око 10 милиона упитника, а добио би око 2 милиона одговора.

Интересантно, истовремено је друга агенција спровела истраживање над мањим скупом људи (4500) и они су имали исправну предикцију.

Пример

Literary Digest је од 1920. спроводио истраживање о победнику Америчких избора. Иако је неколико година био успешан, 1936. резултат истраживања је био погрешан.

Часопис је својим претплатницима слао анонимне упитнике и послао би око 10 милиона упитника, а добио би око 2 милиона одговора.

Интересантно, истовремено је друга агенција спровела истраживање над мањим скупом људи (4500) и они су имали исправну предикцију.

Зашто тако велики узорак је имао лош резултат?

Пристрасност: претплатници часописа су углавном богатији, конзервативни људи који ће вероватно радије гласати за Републиканце.

Шта је добра величина узорка?

Шта је добра величина узорка? Зависи од много фактора:

- величина посматране популације

Шта је добра величина узорка? Зависи од много фактора:

- величина посматране популације
- економска исплативост

Шта је добра величина узорка? Зависи од много фактора:

- величина посматране популације
- економска исплативост
- ниво прецизности који је потребно постићи

Шта је добра величина узорка? Зависи од много фактора:

- величина посматране популације
- економска исплативост
- ниво прецизности који је потребно постићи
- питања која се истражују

Шта је добра величина узорка? Зависи од много фактора:

- величина посматране популације
- економска исплативост
- ниво прецизности који је потребно постићи
- питања која се истражују
- ...

Већина истраживања о мишљењу о некој теми има између 1000 и 2000 испитаника.

Већина истраживања о мишљењу о некој теми има између 1000 и 2000 испитаника.

Прецизност добијена узимањем већег скупа углавном није вредна трошка.

Најбољи начин је **насумично** одабрати јединке у узорку.

Најбољи начин је **насумично** одабрати јединке у узорку.

Ипак у пракси ово је тешко реализовати.

- Стратификација (популација се подели на групе и из сваке групе се бира насумично)

- **Стратификација** (популација се подели на групе и из сваке групе се бира насумично)
- **Кластеровање** (популација се подели на групе и насумично се иабере неколико група)

Бирање узорка – Пример 1

Радио станица се бави истраживањем о легализације марихуане.

Након јављања слушалаца закључак је да је 78% испитаника подржало предлог и да влада одмах треба да донесе закон.

Бирање узорка – Пример 1

Радио станица се бави истраживањем о легализације марихуане.

Након јављања слушалаца закључак је да је 78% испитаника подржало предлог и да влада одмах треба да донесе закон.

Испитаници који су се јавили су само они који слушају ту радио станицу (која можда подржава такав став). Уз то, јавили су се само они испитаници којима је ова тема јако важна.

Бирање узорка – Пример 2

Најчешће се приликом истраживања мишљења о некој теми бирају насумични бројеви телефона. Корисници се позивају и одговарају на питања.

Да ли ипак и овде постоји пристрасност?

Бирање узорка – Пример 2

Најчешће се приликом истраживања мишљења о некој теми бирају насумични бројеви телефона. Корисници се позивају и одговарају на питања.

Да ли ипак и овде постоји пристрасност?

Најсиромашнији људи, бескућници и сличне групе немају телефон.

У добрим истраживањима ће писати колика је маргина грешке.

Ова грешка се управо може десити због проблема са узорком.

У добрим истраживањима ће писати колика је маргина грешке.

Ова грешка се управо може десити због проблема са узорком.

Маргина грешке

Маргина грешке је статистичко изражавање количине случајне грешке приликом узимања узорка.

Обично се дефинише и као радијус интервала поузданости за одређену статистику истраживања.

То је цена коју плаћамо јер не можемо анкетирати целу популацију.

Бирање узорка – Пример 3

У јануару су резултати анкете за неког председничког кандидата показивали да има 54% подршке.

Маргина грешке је 5%.

У јуну, резултати анкете показују да је подршка 56% и новинари закључују да се подршка повећала.

Да ли је закључак исправан?

Бирање узорка – Пример 3

У јануару су резултати анкете за неког председничког кандидата показивали да има 54% подршке.

Маргина грешке је 5%.

У јуну, резултати анкете показују да је подршка 56% и новинари закључују да се подршка повећала.

Да ли је закључак исправан?

Први резултат сугерише да је подршка између 49% и 58%, а други да је подршка између 51% и 61%.

Питања у анкети не смеју да буду пристрасна или вишезначна.

Свако ко одговара на анкету мора на исти начин да разуме питања и да на њих искрено одговори.

- Ипак, ове услове у вези питања није лако постићи.

- Ипак, ове услове у вези питања није лако постићи.
- Пример са *“бројем партнера”*.

- Ипак, ове услове у вези питања није лако постићи.
- Пример са *“бројем партнера”*.
- Пример: *“Да ли читате Политику?”*

- Ипак, ове услове у вези питања није лако постићи.
- Пример са *“бројем партнера”*.
- Пример: *“Да ли читате Политику?”*
- Да ли сваки дан или понекад или једном недељно? Да ли
цео лист или само делове?

- Ипак, ове услове у вези питања није лако постићи.
- Пример са *“бројем партнера”*.
- Пример: *“Да ли читате Политику?”*
- Да ли сваки дан или понекад или једном недељно? Да ли
цео лист или само делове?
- Пример: *“Да ли пијете пуно алкохола?”*

- Ипак, ове услове у вези питања није лако постићи.
- Пример са *“бројем партнера”*.
- Пример: *“Да ли читате Политику?”*
- Да ли сваки дан или понекад или једном недељно? Да ли цео лист или само делове?
- Пример: *“Да ли пијете пуно алкохола?”*
- Шта значи *пуно*? За различите људе то може бити различита вредност.

Добре анкете прво “тестирају/пробају” питања на мањем узорку.

Корелација

Корелација

Корелација је међуоднос или међусобна повезаност између различитих појава представљених вредностима две варијабле. При томе повезаност значи да је вредност једне варијабле могуће с одређеном вероватноћом предвидети на основу сазнања о вредности друге.

Корелација представља и образац варирања варијабли у зависности од начина на који су повезане.

Корелација је често погрешно схваћена.

Корелација је често погрешно схваћена.

Када кажемо да су две променљиве A и B у корелацији, то не значи да међу њима постоји узрочно последична веза.

Корелација је често погрешно схваћена.

Када кажемо да су две променљиве A и B у корелацији, то не значи да међу њима постоји узрочно последична веза.

Наћи узрок неког понашања јако је тешко и предмет је научних истраживања.

A и B су у корелацији може да значи:

- A је узрок појаве B

A и B су у корелацији може да значи:

- A је узрок појаве B
- B је узрок појаве A

A и B су у корелацији може да значи:

- A је узрок појаве B
- B је узрок појаве A
- A и B су случајно повезани без било какве узрочне везе између њих

A и B су у корелацији може да значи:

- A је узрок појаве B
- B је узрок појаве A
- A и B су случајно повезани без било какве узрочне везе између њих
- A и B су обоје зависни од неког трећег фактора C

- Ђаци који пуше цигаре – њихове оцене су лошије

- Ђаци који пуше цигаре – њихове оцене су лошије
- Цена кафе у Орегону – количина кише која пада у Орегону

- Ђаци који пуше цигаре – њихове оцене су лошије
- Цена кафе у Орегону – количина кише која пада у Орегону
- Број димњака у кући – број деце у тој кући

Монаси у Кини: Помрачење месеца изазива небески пас који поједе месец . Зато морају да ударе у велики гонг да би пса отерали.

Монаси у Кини: Помрачење месеца изазива небески пас који поједе месец . Зато морају да ударе у велики гонг да би пса отерали.

Управо погрешно разумевање корелације, узрока и последице води многим **сујеверјима**.

Регресија ка средини и сујеверје

Регресија ка средини

Две вредности које нису у савршеној корелацији, екстремне вредности једне променљиве су у корелацији са неекстремним вредностима друге променљиве, често у корелацији са средњом вредности друге променљиве.

Регресија ка средини

Две вредности које нису у савршеној корелацији, екстремне вредности једне променљиве су у корелацији са неекстремним вредностима друге променљиве, често у корелацији са средњом вредности друге променљиве.

Појам је увео енглески научник **Френсис Галтон** (статистичар, социолог, психолог, антрополог, ...).

Френсис Галтон је испитивао однос висине између очева и синова:

- Открио је очигледно: високи очеви имају високе синове, а ниски очеви имају ниске синове.

Френсис Галтон је испитивао однос висине између очева и синова:

- Открио је очигледно: високи очеви имају високе синове, а ниски очеви имају ниске синове.
- Међутим, открио је још нешто изненађујуће. . .

Френсис Галтон је испитивао однос висине између очева и синова:

- Открио је очигледно: високи очеви имају високе синове, а ниски очеви имају ниске синове.
- Међутим, открио је још нешто изненађујуће. . .
- Веома високи очеви имају синове који су нижи од њих.

Френсис Галтон је испитивао однос висине између очева и синова:

- Открио је очигледно: високи очеви имају високе синове, а ниски очеви имају ниске синове.
- Међутим, открио је још нешто изненађујуће. . .
- Веома високи очеви имају синове који су нижи од њих.
- Веома ниски очеви имају синове који су виши од њих.

Френсис Галтон је испитивао однос висине између очева и синова:

- Открио је очигледно: високи очеви имају високе синове, а ниски очеви имају ниске синове.
- Међутим, открио је још нешто изненађујуће. . .
- Веома високи очеви имају синове који су нижи од њих.
- Веома ниски очеви имају синове који су виши од њих.
- Шта ово значи?

- Ово је пример несавршене корелације две променљиве.

- Ово је пример **несавршене корелације** две променљиве.
- Много тога утиче на висину деце.

- Ово је пример **несавршене корелације** две променљиве.
- Много тога утиче на висину деце.
- Мајчина висина.

- Ово је пример **несавршене корелације** две променљиве.
- Много тога утиче на висину деце.
- Мајчина висина.
- Начин одгајања, исхране, место одгајања, бављење различитим активностима итд. . .

- Ово је пример **несавршене корелације** две променљиве.
- Много тога утиче на висину деце.
- Мајчина висина.
- Начин одгајања, исхране, место одгајања, бављење различитим активностима итд. . .
- Пуно фактора мора да се поклопи да би особа била екстремно висока или ниска.

- Вероватноћа да се сви услови поклопе је јако мала.

- Вероватноћа да се сви услови поклопе је јако мала.
- То објашњава зашто случај екстремно високог оца је у корелацији са неекстремним вредностима висине сина.

- Вероватноћа да се сви услови поклопе је јако мала.
- То објашњава зашто случај екстремно високог оца је у корелацији са неекстремним вредностима висине сина.
- Тј. ова појава је предвидива и назива се регресија ка средини.

Наводно врхунски спортисти су сујеверни и јако се боје да се појаве на насловној страни часописа *Sports Illustrated*. Након појаве у часопису њихови резултати су лошији.

Регресија ка средини – пример са спортистима

Наводно врхунски спортисти су сујеверни и јако се боје да се појаве на насловној страни часописа *Sports Illustrated*. Након појаве у часопису њихови резултати су лошији.

Баксуз је појавити се на насловној страни или регресија ка средини?

- Они су позвани на насловну страну јер су имали **изванредан, несвакидашњи** успех.

Регресија ка средини – пример са спортистима

- Они су позвани на насловну страну јер су имали **изванредан, несвакидашњи** успех.
- Такав успех је често сплет разних фактора: веома напорног тренирања, изузетног здравственог стања, повољних услова такмичења, помало среће,...

- Они су позвани на насловну страну јер су имали **изванредан, несвакидашњи** успех.
- Такав успех је често сплет разних фактора: веома напорног тренирања, изузетног здравственог стања, повољних услова такмичења, помало среће, . . .
- **Ово је екстремна појава која се дешава када се много фактора поклопи, а за све то је вероватноћа јако мала.**

Регресија ка средини – пример са спортистима

- Они су позвани на насловну страну јер су имали **изванредан, несвакидашњи** успех.
- Такав успех је често сплет разних фактора: веома напорног тренирања, изузетног здравственог стања, повољних услова такмичења, помало среће, . . .
- **Ово је екстремна појава која се дешава када се много фактора поклопи, а за све то је вероватноћа јако мала.**
- Наравно да ће у будућности њихови резултати бити лошији.

Графици и илустрације: вреде
као хиљаду лажи

Марк Твен (1835 – 1910)

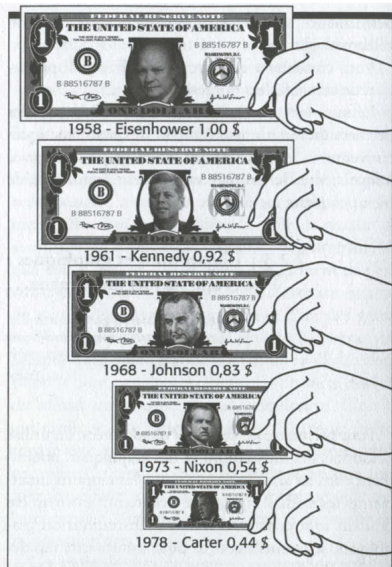
Прво сазнај све чињенице, а онда их можеш изменити како год желиш.

Често користимо графике и илустрације да визуелно прикажемо податке.

На овај начин информације се лакше и брже преносе и често је податке лакше разумети.

Посебно се овакав вид приказа информација користи у научним чланцима, финансијским извештајима и медијским извештавањима.

Графици и илустрације – вредност долара у новинском извештавању



- Вредност долара између 1958. и 1978. године је опала.
- Године 1978. је било потребно 2\$ за оно што је 1958. године коштало 1\$ долар.
- Вредност долара је била **дупло мања**.

Графици и илустрације – вредност долара у новинском извештавању



Ипак, сликар је приказао тако да је вредност **четири пута мања**.

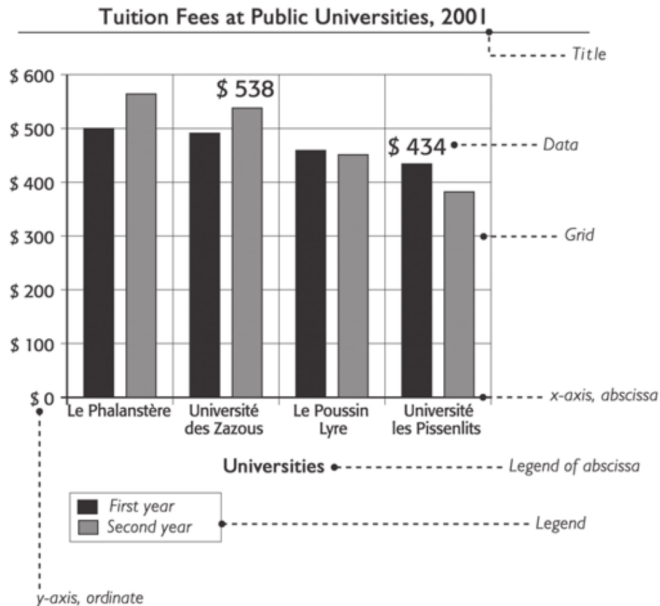
Едвард Туфте (1942 –), амерички статистичар

Задаје принцип: Репрезентација бројева које се физички могу измерити на посматраној слици мора да буде директно пропорционална вредностима које репрезентује.

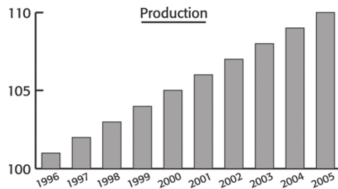
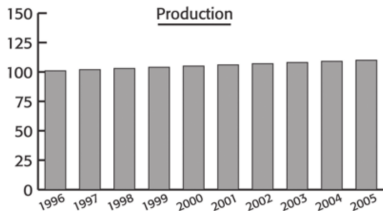
Свако одступање од овог принципа представља лаж.

Критичар начина на који се користи *Power Point*.

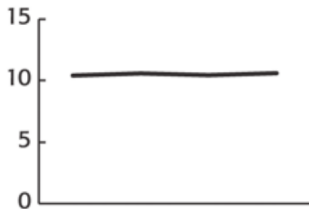
Графици и илустрације – Пример добре табеле



Графици и илустрације – Варање са табелама, промене на Y оси

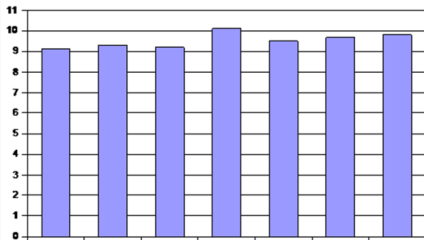


Графици и илустрације – Варање са табелама, промене на Y оси

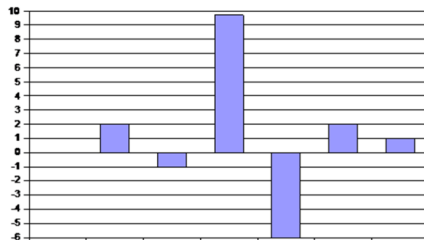


Графици и илустрације – Варање са табелама, промене на Y оси

Temperatura u C



Promena temperature u %



Проценти и подвале

Цена му се смањила за 250%.

Цене ћемо смањити 100%.

Сваки прозвод снижен за 10%. Уколико купите 3 производа, снижење је 30%.

Укупна цена: $a + b + c = 100$ динара.

Снижено за 30% је 70 динара.

$$0.9a + 0.9b + 0.9c = 0.9(a + b + c) = 0.9 \cdot 100 = 90$$

Значи није укупно 30%, већ 10%.

Проценти се не сабирају.

Аутомобил је на првом месту узрока несрећних случајева.

Аутомобил је на првом месту узрока несрећних случајева.

Зависи од категоризације несрећних случајева:

- **Тачно ако:** 40% аутомобил, 20% оклизнуће се у купатилу, 20% попити случајно отров, 20% задавити се сувим ђевреком

Аутомобил је на првом месту узрока несрећних случајева.

Зависи од категоризације несрећних случајева:

- **Тачно ако:** 40% аутомобил, 20% оклизнуће се у купатилу, 20% попити случајно отров, 20% задавити се сувим ђевреком
- **Нетачно ако:** 40% аутомобил, 60% незгоде у кући

После јучерашњег повећања акција за 20%, данас су доживеле 2 пада од 9%.

После јучерашњег повећања акција за 20%, данас су доживеле 2 пада од 9%.

Да ли имамо повећање од 2%?

После јучерашњег повећања акција за 20%, данас су доживеле 2 пада од 9%.

Да ли имамо повећање од 2%?

вредност акција: x

После јучерашњег повећања акција за 20%, данас су доживеле 2 пада од 9%.

Да ли имамо повећање од 2%?

вредност акција: x

повећање 20%: $y = x + 0.2x = 1.2x$

После јучерашњег повећања акција за 20%, данас су доживеле 2 пада од 9%.

Да ли имамо повећање од 2%?

вредност акција: x

повећање 20%: $y = x + 0.2x = 1.2x$

прво смањење 9%: $y_1 = y - 0.09y = 1.2x - 0.09(1.2x) = 1.092x$

После јучерашњег повећања акција за 20%, данас су доживеле 2 пада од 9%.

Да ли имамо повећање од 2%?

вредност акција: x

повећање 20%: $y = x + 0.2x = 1.2x$

прво смањење 9%: $y_1 = y - 0.09y = 1.2x - 0.09(1.2x) = 1.092x$

друго смањење 9%:

$y_2 = y_1 - 0.09y_1 = 1.092x - 0.09(1.092x) = 0,99372x$

После јучерашњег повећања акција за 20%, данас су доживеле 2 пада од 9%.

Да ли имамо повећање од 2%?

вредност акција: x

повећање 20%: $y = x + 0.2x = 1.2x$

прво смањење 9%: $y_1 = y - 0.09y = 1.2x - 0.09(1.2x) = 1.092x$

друго смањење 9%:

$y_2 = y_1 - 0.09y_1 = 1.092x - 0.09(1.092x) = 0,99372x$

Није било повећања, већ крајњи резултат показује смањење укупне вредности.



Normand Baillargeon.

A Short Course in Intellectual Self Defense: Find Your Inner Chomsky.

Seven Stories Press, 2011.