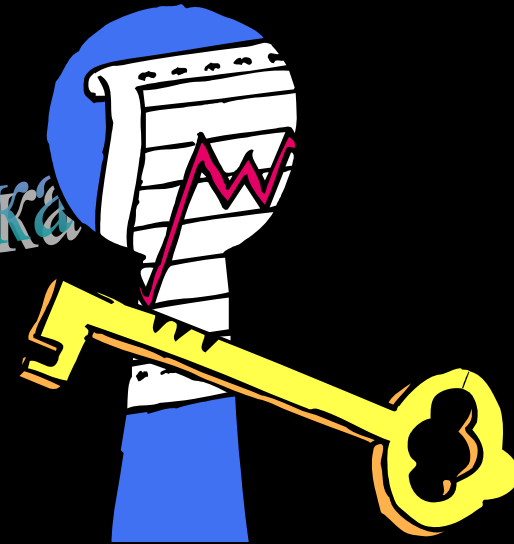


Статистиката е наука за

- събиране,
- организиране,
- обобщаване,
- анализиране, и
- интерпретиране

Що е статистика

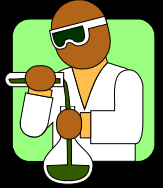


на данни с цел да се вземе по-ефективно решение.

Основни елементи на статистиката

- Събиране на данни
- Обобщаване на данни
- Интерпретиране на данни
- Вземане на решения от данни

Събиране на данни



- Определяне предмета на изследване
 - Редуцира ли аспирин риска от сърдечен инфаркт?
- Наблюдения
 - Наблюдения на хора, вземащи аспирин и не вземащи аспирин



Време за някой дефиниции

Популация: Групата, която притежава изучаваната характеристика

Индивид : всеки член на популацията

Извадка - подмножество на популацията

Броят **n** на елементите на извадката се нарича **обем** на извадката.

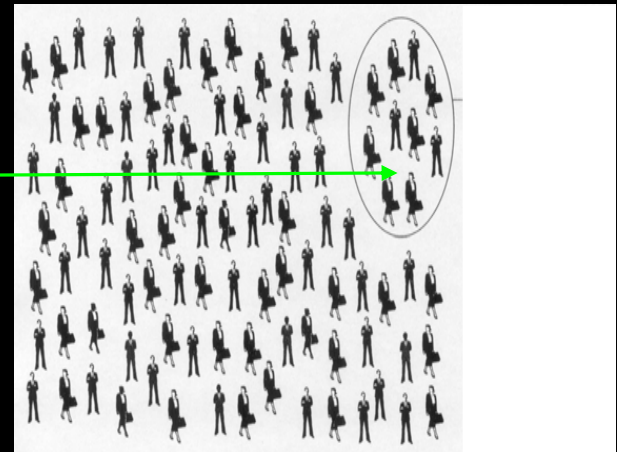
Пример: Изучаване въздействието на аспирина

Популация: всички хора от дадена възрастова група

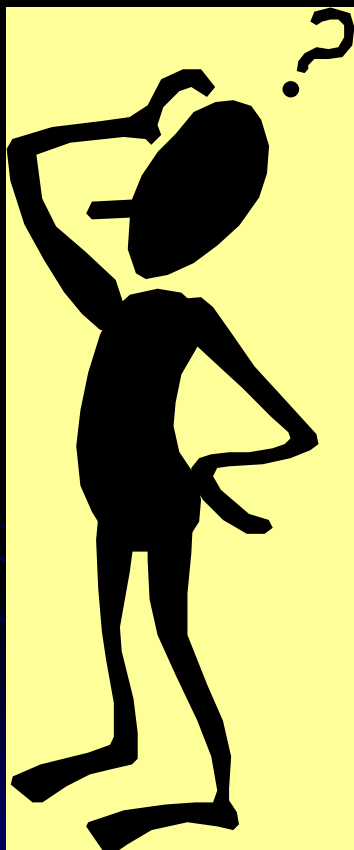
Индивид : всеки човек

Извадка : избират се само 100 човека и се наблюдават

100= обем на извадката



Защо извадка?



В повечето изследвания
в е трудно да се получи
информация от цялата
популация. Тогава на
основата на извадката
ние оценяваме или
правим изводи за цялата
популация.

Още дефиниции

променливи

– изследваната характеристика(и)
на индивидите в популацията

данни

- Списък от стойности от наблюденията
на изследваната характеристика (и)

Пример

Изследване на средния успех на студентите в ПУ

Променлива: среден успех

Данни: 3,82; 4,95; 4,82; 3,49; 5,70

Видове променливи

Качествени променливи — изследваната характеристика не се измерва числено

При наблюдението им се определят категории

Количествени променливи — измерва се числено

Дискретни
променливи

Непрекъснати
променливи

Примери

Сини
Кафяви
Черни и пр.

- цвят на очи
- пол,
- Мнение относно преподавател

Жена
Мъж

Много добро
Добро
Лошо
Нямам мнение

- баланс в банкова сметка
- ръст
- успех
- възраст

**Качествени
променливи**

**Количествени
променливи**

**Дискретни
променливи**

**Непрекъснати
променливи**

**Качествени
данни**

**Количествени
данни**

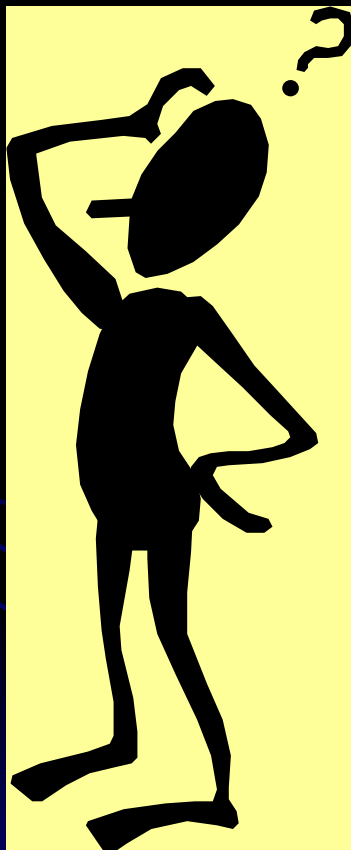
**Дискретни
данни**

**Непрекъснати
данни**

Само определени стойности
са възможни (има скок между
възможните стойности)

Теоретично, всяка стойност
в интервал е възможна

Защо е важен видът на данните?



Видът на данните
определя и
статистическия анализ,
който ще се използва

Как се събират данни?

Има различни методи, но ние ще разгледаме само **случайни извадки**.

С разглеждането на случайните извадки, ще считаме, че тя представлява адекватно популацията, което ни дава основание да считаме, че заключенията, направени от извадката са верни и за популацията.

Разбира се при това има някаква несигурност—вероятност за грешка!!!

Организиране и обобщаване на данните

Графично представяне на данните

- Зависи от типа данни
- Зависи от това, какво искаме да илюстрираме
- Зависи от статистическия софтуер, с който разполагаме

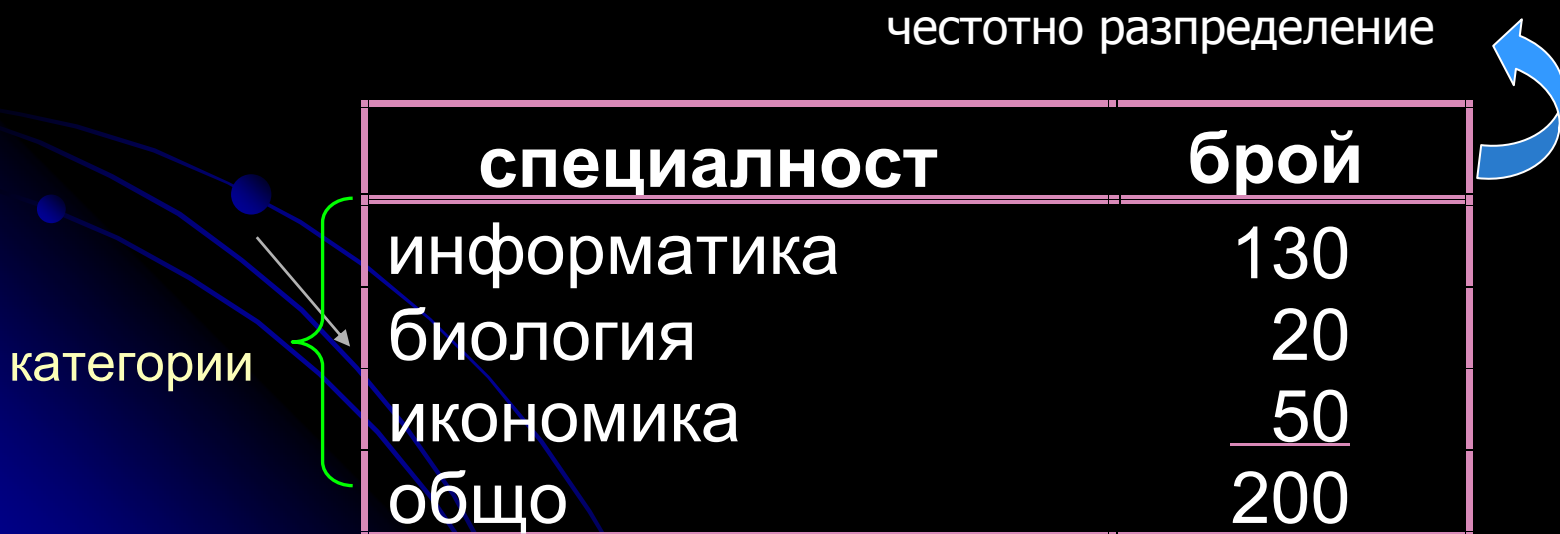
Качествени данни

Избрани са по случаен начин 200 първокурсници в ПУ и е записана тяхната специалност

Пресмятат се ЧЕСТОТИ и се оформя

честотно разпределение

категории



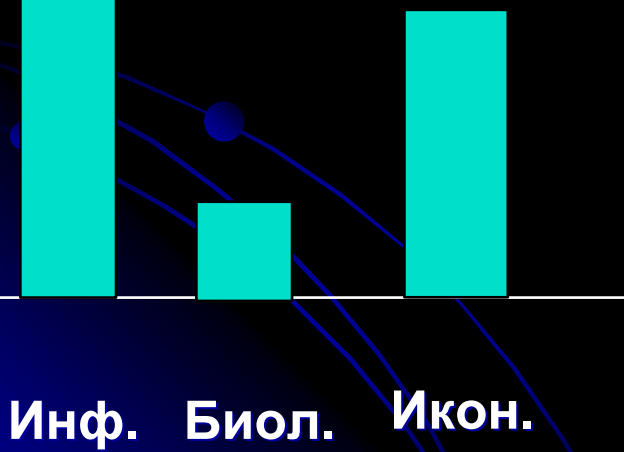
| специалност | брой |
|-------------|------|
| информатика | 130 |
| биология | 20 |
| икономика | 50 |
| общо | 200 |

Представяне графично на събраните данни:

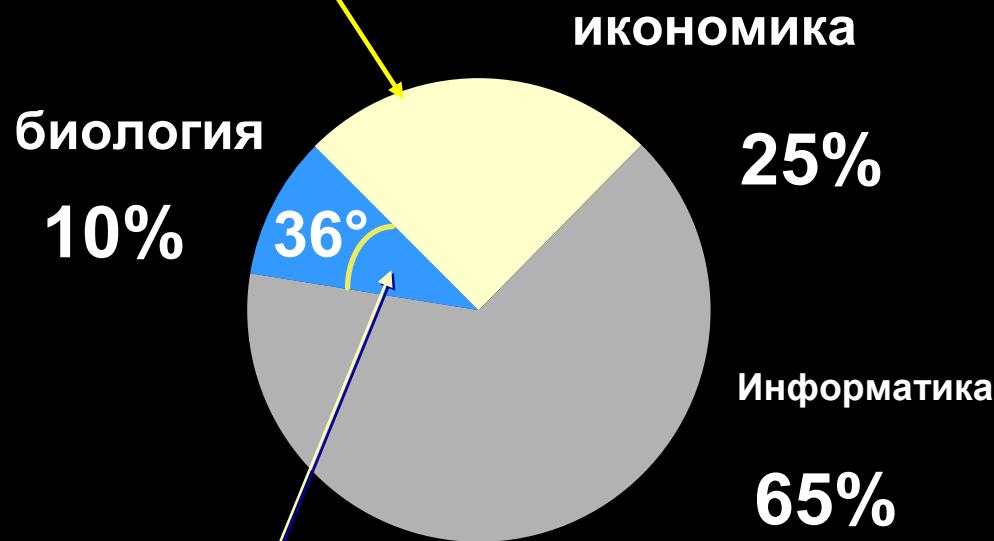
Специалности

Хистограма

Всяка категория се нанася на хоризонталната ос и се чертаят правоъгълници, чиято височина е равна на честотата



Кръг се разделя на сектори, като всеки представя различни категории. Лицето на всеки сектор е пропорционален на честотата.



$$(360^\circ) (10\%) = 36^\circ$$

Количествени данни

Дискретни данни

Графичното представяне с хистограма е подобно на качествените данни.

Непрекъснати данни

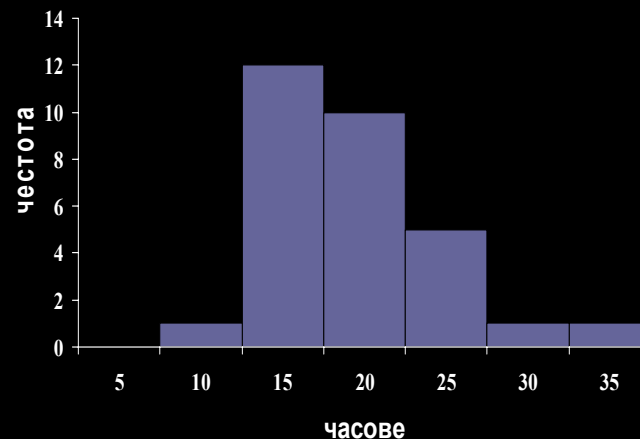
Първо, данните се групират, като получаваме честотно разпределение

ПРИМЕР: Направена е случайна извадка от 30 студенти, които са запитани за броя часове, прекарани в учене през последната седмица:

| | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|------|
| 15,0 | 23,7 | 19,7 | 15,4 | 18,3 | 23,0 | 14,2 | 20,8 | 13,5 | 20,7 | 17,4 |
| 18,6 | 12,9 | 20,3 | 13,7 | 21,4 | 18,3 | 29,8 | 17,1 | 18,9 | 10,3 | 26,1 |
| 15,7 | 14,0 | 17,8 | 33,8 | 23,2 | 12,9 | 27,1 | 16,6 | | | |

| Часове | честота | Относителна честота |
|--------------|---------|---------------------|
| 7,5 до 12,5 | 1 | $1/30=0,0333$ |
| 12,5 до 17,5 | 12 | $12/30=0,400$ |
| 17,5 до 22,5 | 10 | $10/30=0,333$ |
| 22,5 до 27,5 | 5 | $5/30=0,1667$ |
| 27,5 до 32,5 | 1 | $1/30=0,0333$ |
| 32,5 до 37,5 | 1 | $1/30=0,0333$ |
| ОБЩО | 30 | $30/30=1$ |

хистограма



Защо различни методи?



Качествените и
количествените данни
имат съвсем различно
поведение и затова се
изучават по различен
начин.

Описателна статистика

Какво можем да опишем?

Какво е “местоположението” или
“центъра” на данните?

Как варират данните?

Мерки за “местоположението” на данните



Средна стойност

Медиана

Мода

Средна стойност

- Ако описва популацията, се нарича популационна средна стойност и се означава с μ (**параметър**)
- Ако описва извадката се нарича извадково средно и се означава с \bar{x} (**статистика**)
- Използва се само за количествени данни.
- Съществено се влияе от всички данни.

Формула

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$x_1, x_2, x_3, \dots, x_n$

Стойностите от данните

Пример: Случайно избрани студенти, взели даден тест, са получили следните точки **14, 15, 17, 16, 15**

Тогава средният им брой точки е извадково средно

$$\bar{X} = \frac{\sum X}{n} = \frac{14 + \dots + 15}{5} = \frac{77}{5} = 15,4$$

Медиана

Данните са подредени във **възходящ ред**.

Медианата е точка, за която 50% от данните са по-малки от нея.

- Данните, по-малки от медианата са точно толкова, колкото и данните по-големи от нея.
- Използва се само за количествени данни.
- При нечетен брой данни, медианата е =средния елемент на данните
- При четен брой данни, медианата е аритметично средното на двата средни елемента;

Примери

6,72 3,46 3,60 6,44 26,70

3,46 3,60 6,44 6,72 26,70 (наредени данни)

(нечетен брой данни)

Има среда

Медианата е 6,44

6,72 3,46 3,60 6,44

3,46 3,60 6,44 6,72 (наредени данни)

(четен брой данни)

3,60 + 6,44

2

Медианата е 5,02

Мода

Модата е най-често срещаната стойност в данните .

- Данните могат да имат повече от една мода.
- Подходяща е за всеки вид данни, но най-често се използва при качествени данни или дискретни данни с малък брой възможни стойности

а. 5 5 5 3 1 5 1 4 3 5

Модата е 5

б. 1 2 2 2 3 4 5 6 6 6 7 9

Бимодална : две моди 2 и 6

в. 1 2 3 6 7 8 9 10

Няма мода

Мерки за “разсейването” на данните

Тези мерки са подходящи само за количествени данни.

Дисперсия

- Ако описва популацията, се нарича популационна дисперсия и се означава с σ^2 (**параметър**)
- Ако описва извадката се нарича извадкова дисперсия и се означава с s^2 (**статистика**)
- Използва се само за количествени данни.
- Съществено се влияе от всички данни.

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

Стандартно отклонение

- Ако описва популацията, се нарича популационна дисперсия и се означава с σ (**параметър**)
- Ако описва извадката се нарича извадкова дисперсия и се означава с s (**статистика**)
- Мерните единици са същите както и мерните единици на данните.
- Измерва отклонението на данните от тяхната средна стойност.
- Съществено зависи от всички данни.

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

Пример

Фирма понякога наема носачи за почасова работа, като им плаща в зависимост от предлагането. Случайно са избрани 5 човека, работили почасово в тази фирма, на които се е оказало, че фирмата е плащала по

7 лв, 5 лв, 11 лв, 8лв, 6 лв на час.

Намерете стандартното отклонение .

$$\bar{X} = \frac{\sum X}{n} = \frac{37}{5} = 7,40$$

$$\begin{aligned} s^2 &= \frac{\sum (X - \bar{X})^2}{n - 1} = \frac{(7 - 7,4)^2 + \dots + (6 - 7,4)^2}{5 - 1} \\ &= \frac{21,2}{5 - 1} = 5,30 \end{aligned}$$

$$s = \sqrt{s^2} = \sqrt{5,30} = 2,30$$

Извадково стандартно отклонение

Важно!!!

Извадковите характеристики

- Извадково средно
- Извадкова дисперсия (изв. станд. откл.)

зависят от извадката =>

те имат поведение на случайни величини.

Като случайни величини те си имат разпределение.

Разпределението на извадковото средно

Нека $\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$

Разглеждаме частен случай:

Нека случайната извадка с обем n е от нормално разпределена популация със средна стойност μ и дисперия σ^2 , т.е. $N(\mu, \sigma^2)$

$$EX_i = \mu$$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$\text{Дисп.} X_i = \sigma^2$$

$$E\bar{X} = \frac{EX_1 + EX_2 + \dots + EX_n}{n} = \frac{\mu + \mu + \dots + \mu}{n} = \mu$$

Средна стойност на извадковото средно

$$\text{Дисп}\bar{X} = \frac{\text{Дисп.} X_1 + \text{Дисп} X_2 + \dots + \text{Дисп} X_n}{n^2} = \frac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Дисперсия на извадковото средно

Извадковото средно е нормално разпределено, т.е.

$$\bar{X} \in N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \in N(0,1)$$

Ако популацията не е нормално разпределена, то съгласно ЦГТ

Извадка с обем n е направена от алтернативна популация- всеки елемент притежава или не притежава дадена характеристика.

Нека X = брой индивиди от извадката, които притежават характеристиката

статистика

$$\hat{p} = \frac{x}{n}$$

Разглеждаме опити на Бернули – n опита и $p = P(\text{Успех}) = P(\text{отделния елемент да притежава характеристиката})$

$X = \{\text{брой } \textit{Успехи} \text{ в тези } n \text{ опита}\}$

биномно разпределение

Стандартно
нормално

При голямо n

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)}} \sqrt{n}$$

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

Статистиката \hat{p} има средна стойност p и дисперсия

$$\text{Дисп. } \hat{p} = \frac{p(1-p)}{n}$$