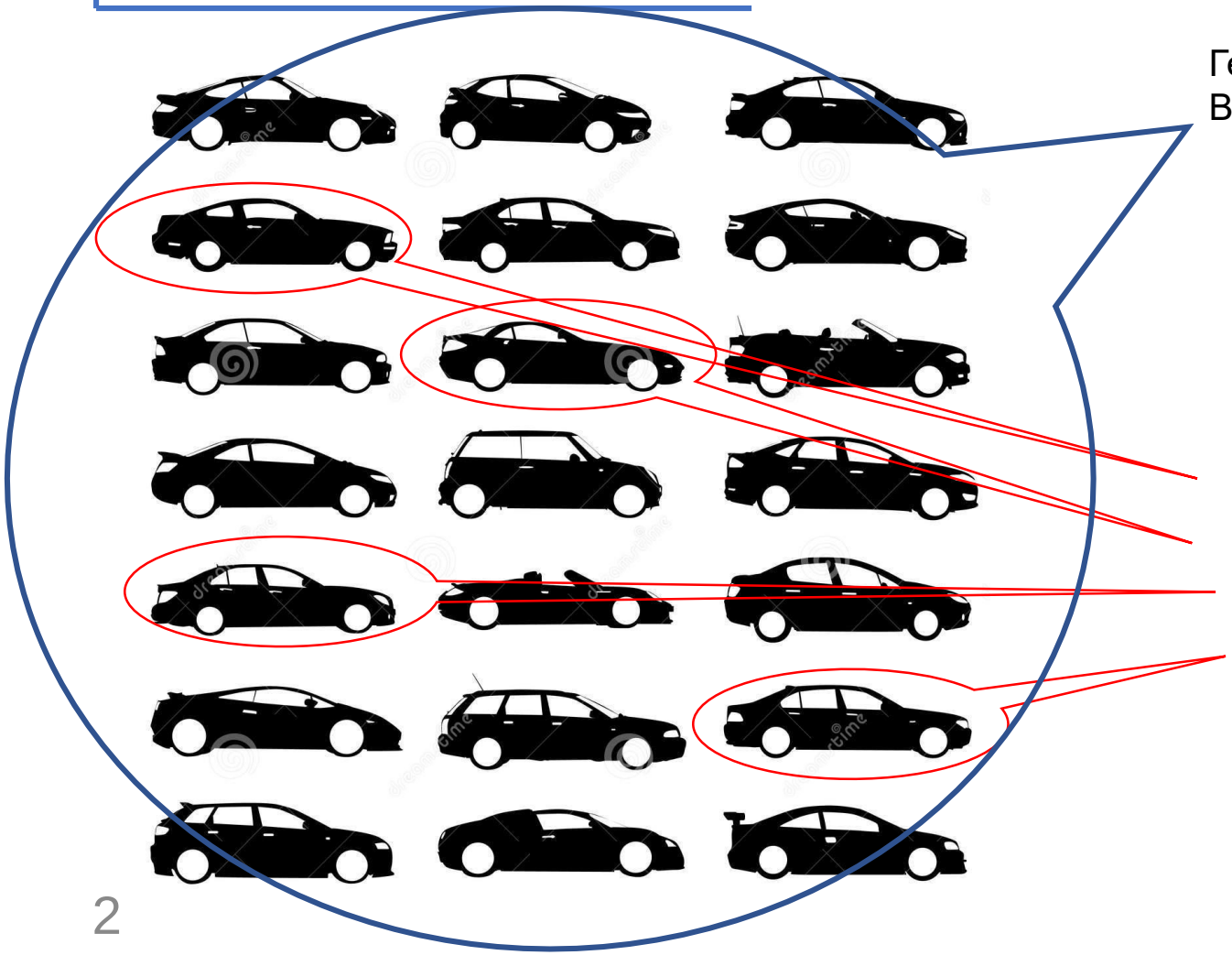
ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР *МГТУ им. Н. Э. Баумана*

Data Science
Описательная статистика

Корпоративное обучение на базе
Образовательного центра МГТУ им. Н. Э. Баумана
под управлением МИЦ «Композиты России»

Докладчик: Панфилов И.А. канд. техн. наук, доцент

Генеральная совокупность объектов и выборка



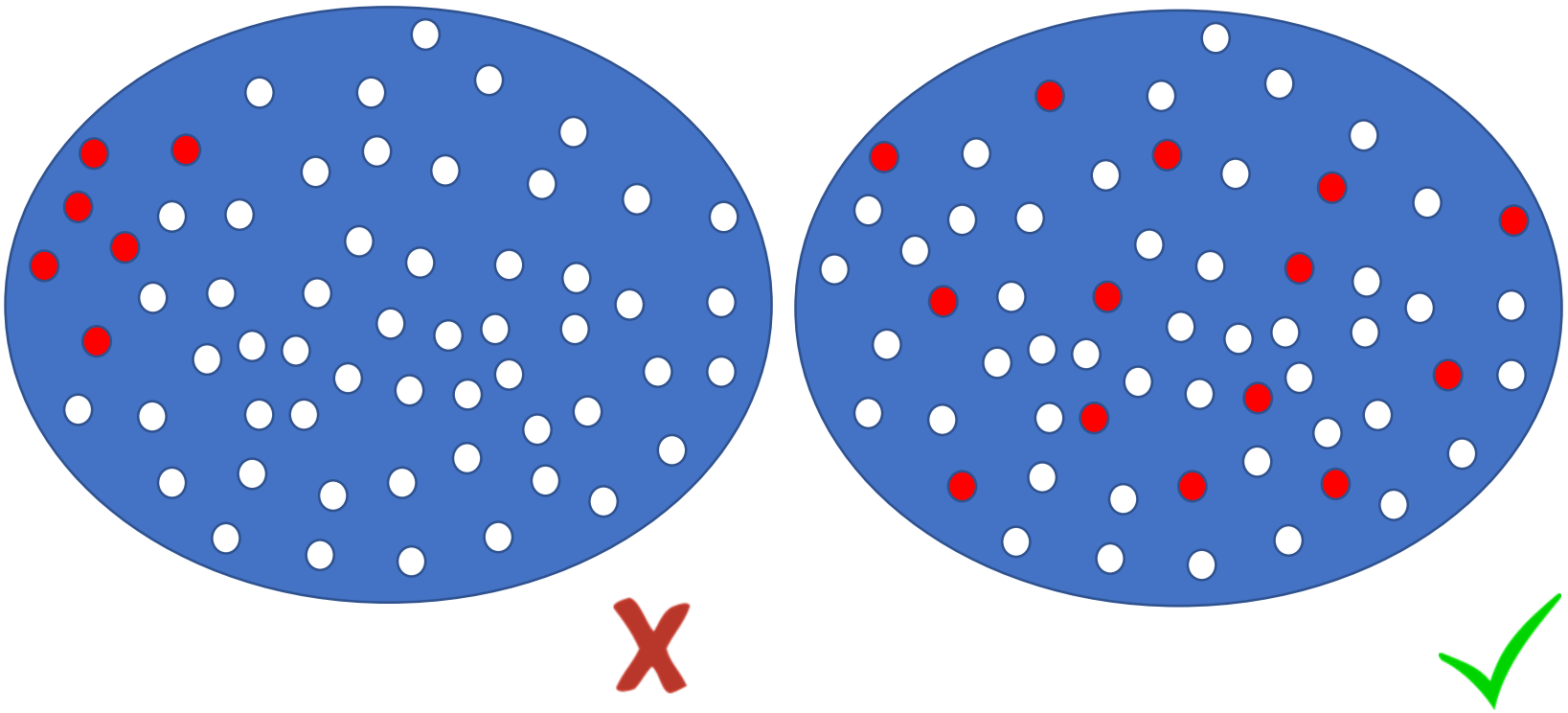
Генеральная совокупность:
ВСЕ автомобили в городе

Выборка:
5% от ВСЕХ автомобилей

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

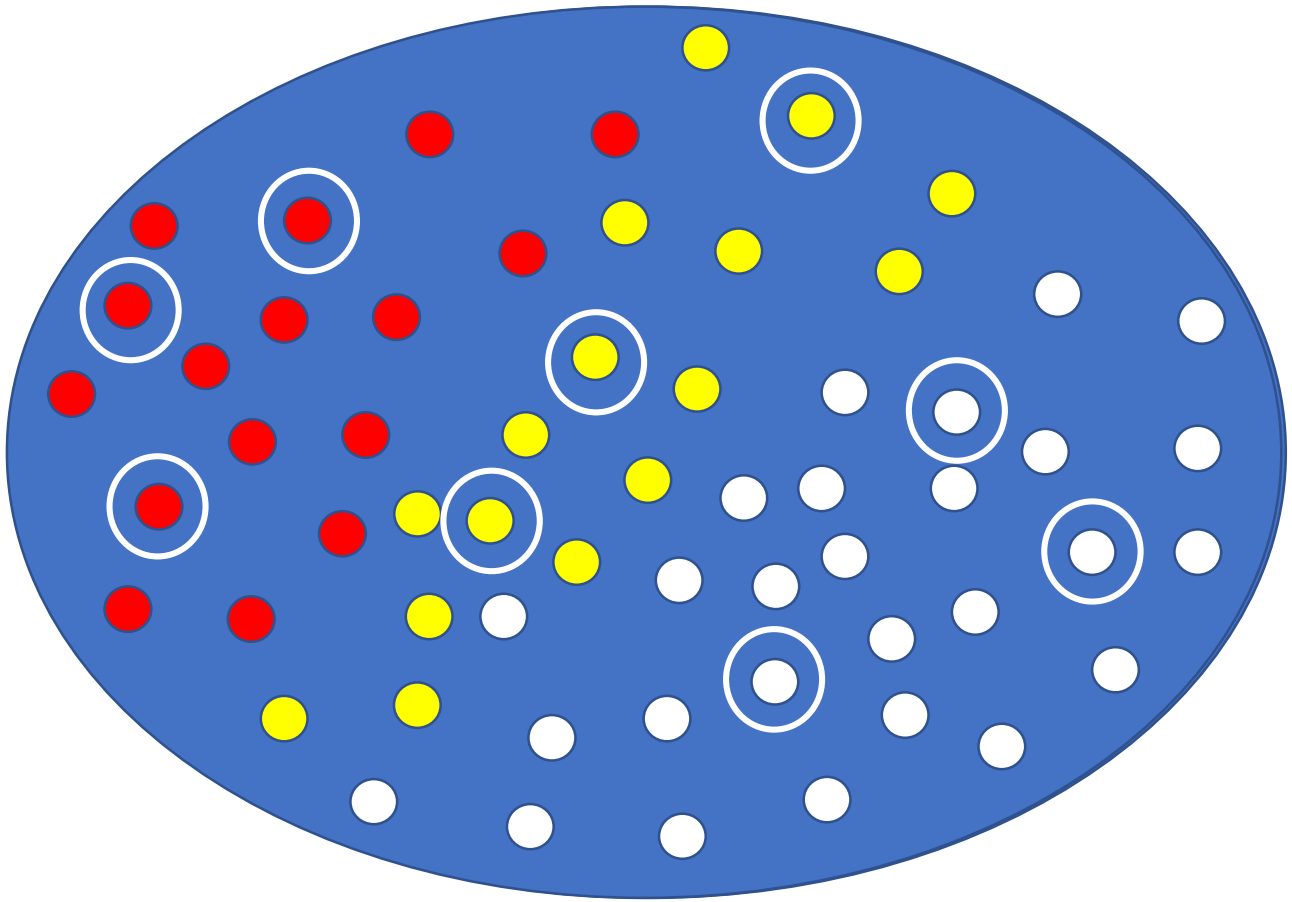
Репрезентативная выборка

простая случайная выборка (simple random sample)



ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

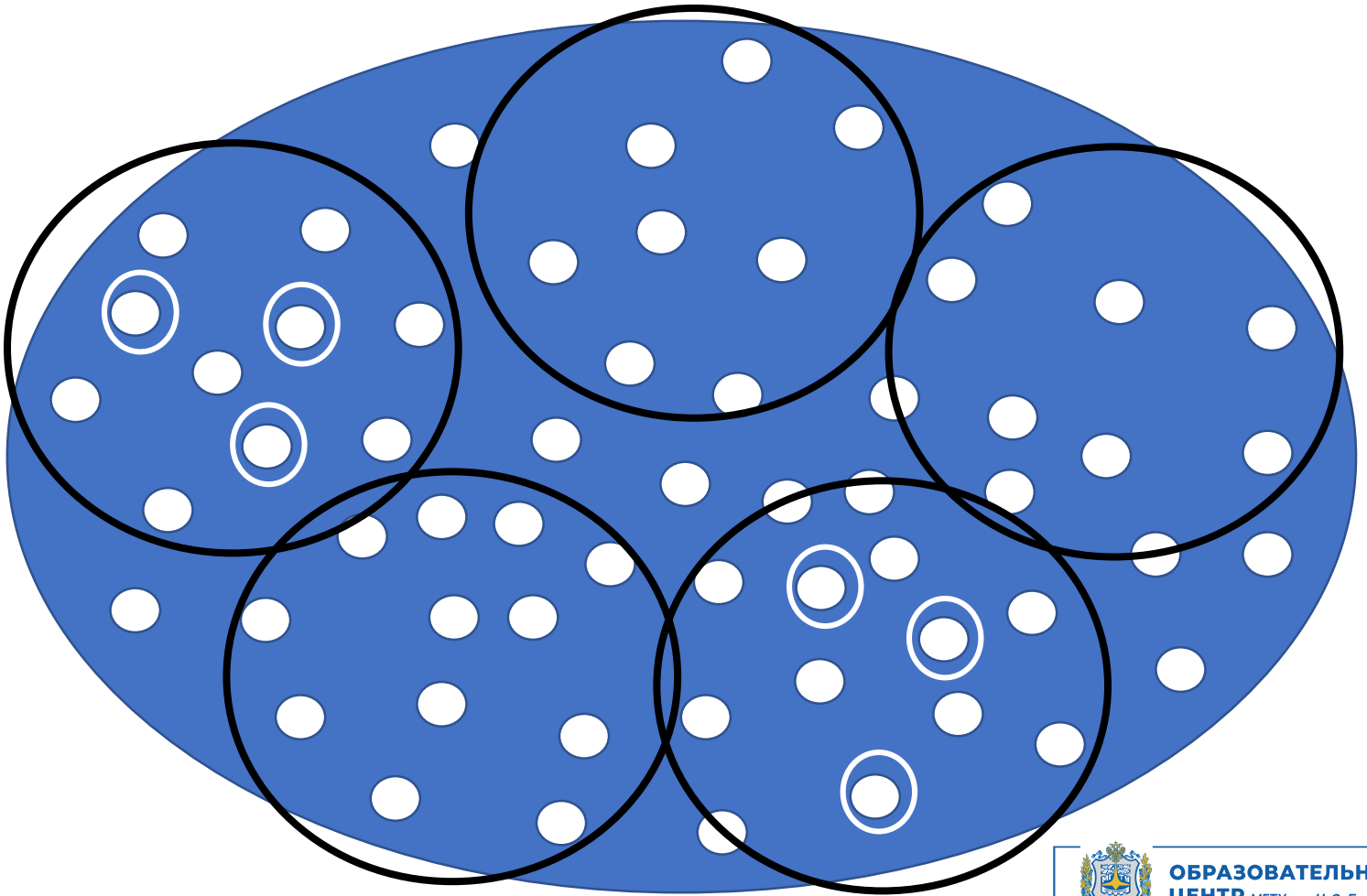
Стратифицированная выборка
(stratified sample)



ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

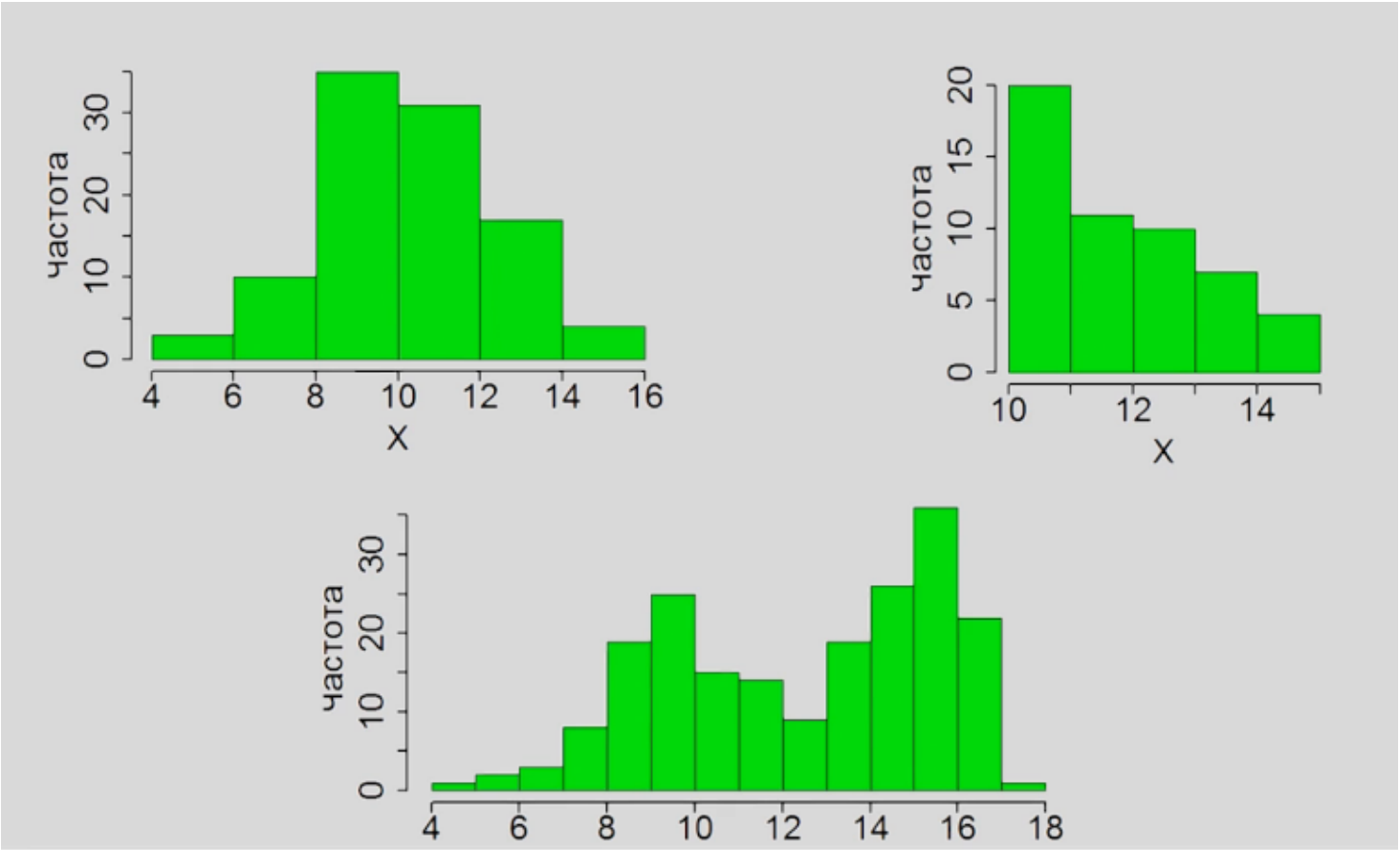
Групповая выборка

cluster sample



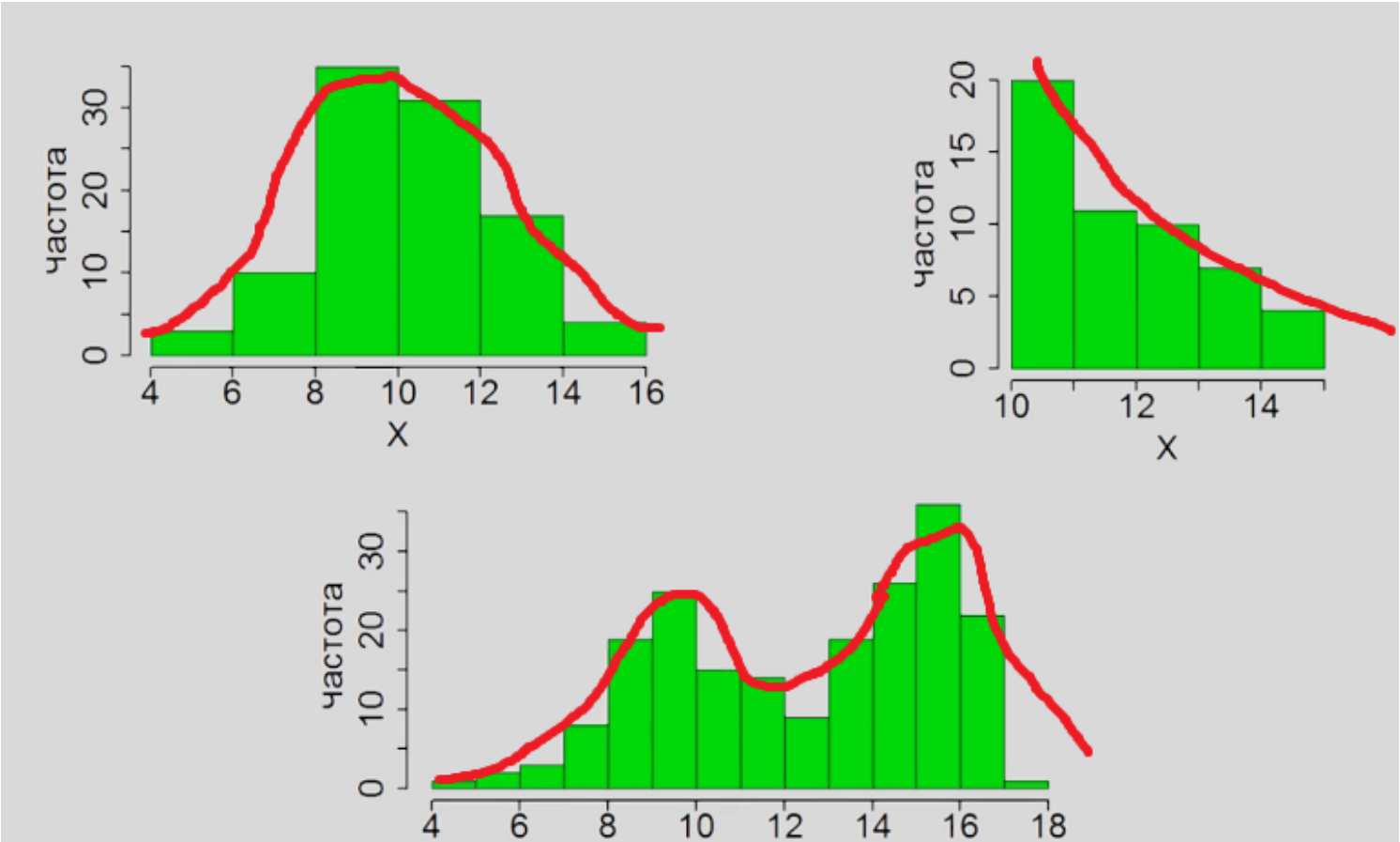
ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Гистограмма частот



ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Характер распределения

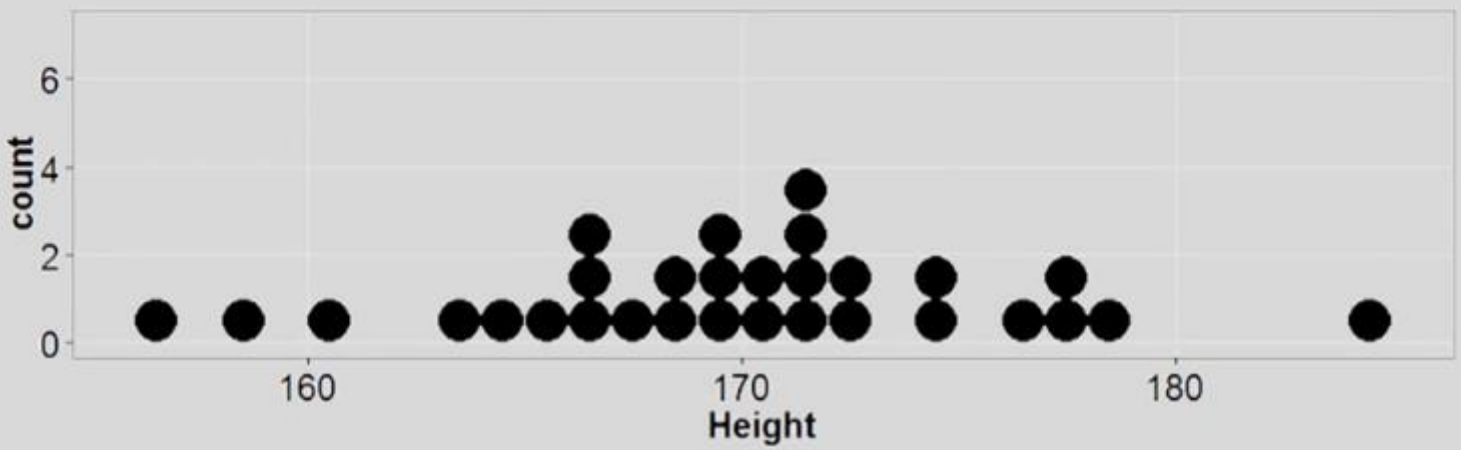


ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Меры центральной тенденции

Мода (Mode) – значение измеряемого признака, которое встречается максимально часто.

185 175 170 169 171 172 175 157 170 172 167 173 168 167 166
167 169 172 177 178 165 161 179 159 164 178 172 170 173 171



ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

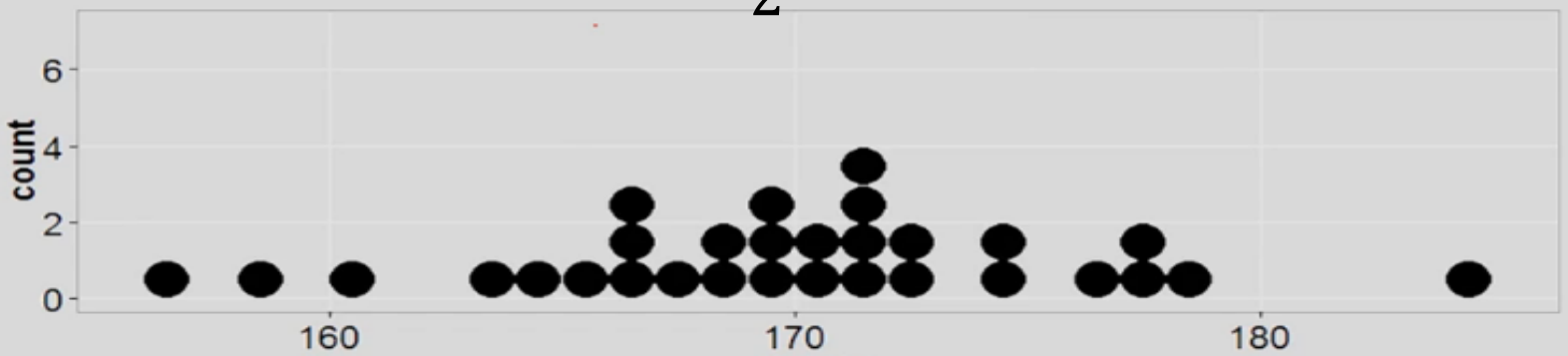
Медиана

Медиана (median) – значение признака, которое делит упорядоченное множество данных пополам.

157 159 161 164 165 166 167 167 167

157 159 161 164 165 166 167 167 167 168 169 169 170 170 170
171 171 172 172 172 172 173 173 175 175 177 178 178 179 185

$$Mc = \frac{170 + 171}{2} = 170{,}5$$



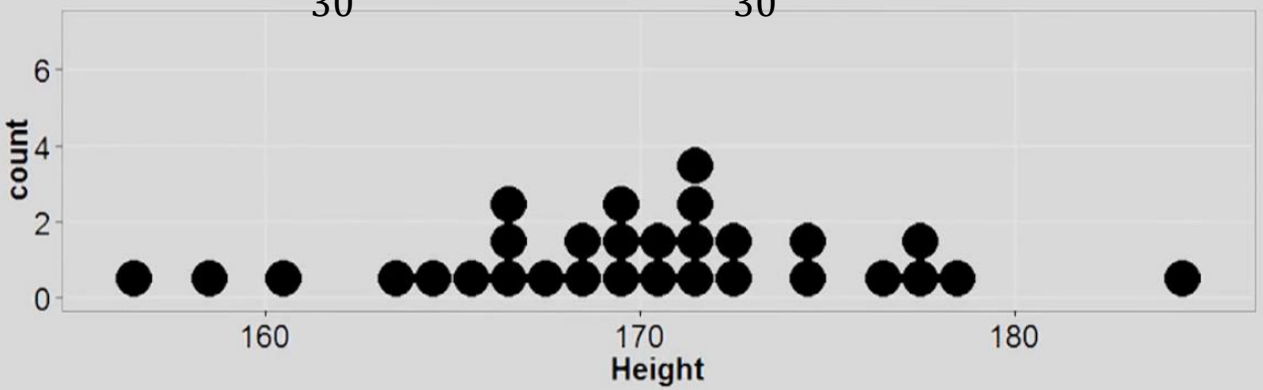
ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Среднее значение

Среднее значение (mean, среднее арифметическое) сумма всех значений измеренного признака, деленная на количество измеренных значений.

185 175 170 169 171 172 175 157 170 172 167 173 168 167 166
167 169 172 177 178 165 161 179 159 164 178 172 170 173 171

$$\bar{x} = \frac{x_1 + x_1 + \cdots x_1}{30} = \frac{185 + 175 + 170 + \cdots + 171}{30} = 170{,}4$$



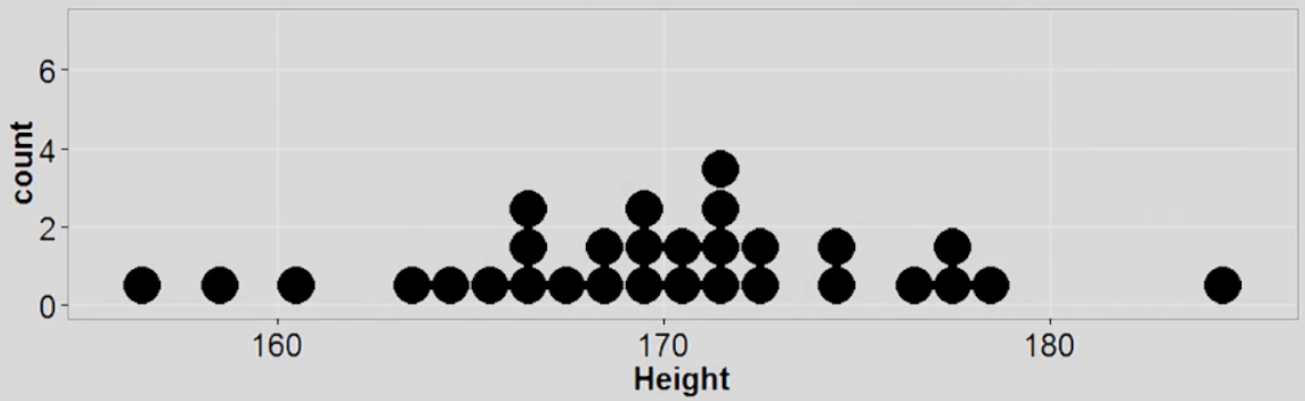
ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Мера изменчивости

размах (range)

Размах (Range) - разность максимального и минимального значения.

$$R = X_{max} - X_{min}$$



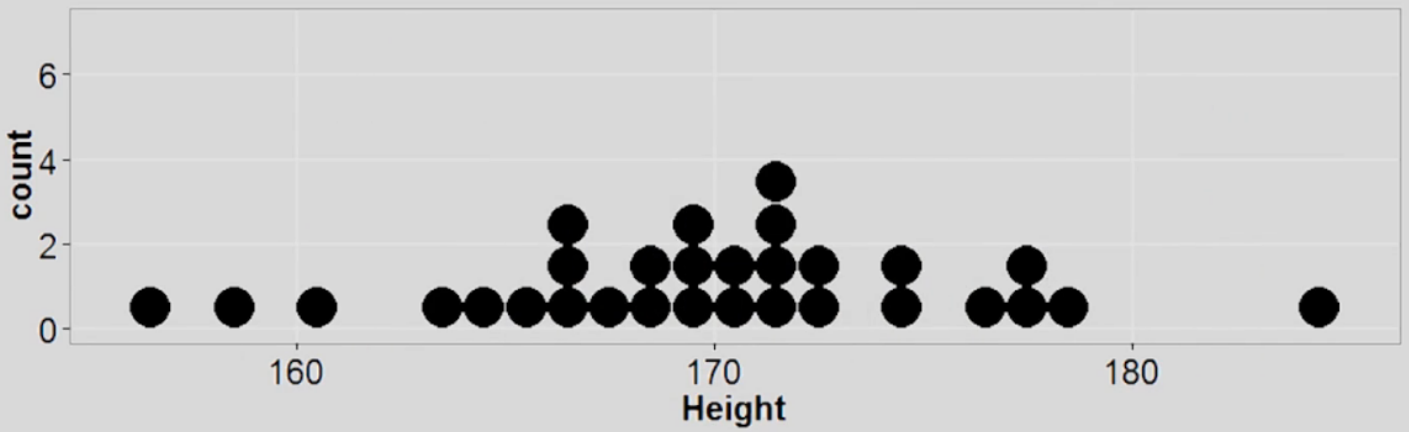
ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Мера изменчивости

дисперсия (variance), среднеквадратическое отклонение (standard deviation)

Дисперсия (variance) – средний квадрат отклонений индивидуальных значений признака от их средней величины.

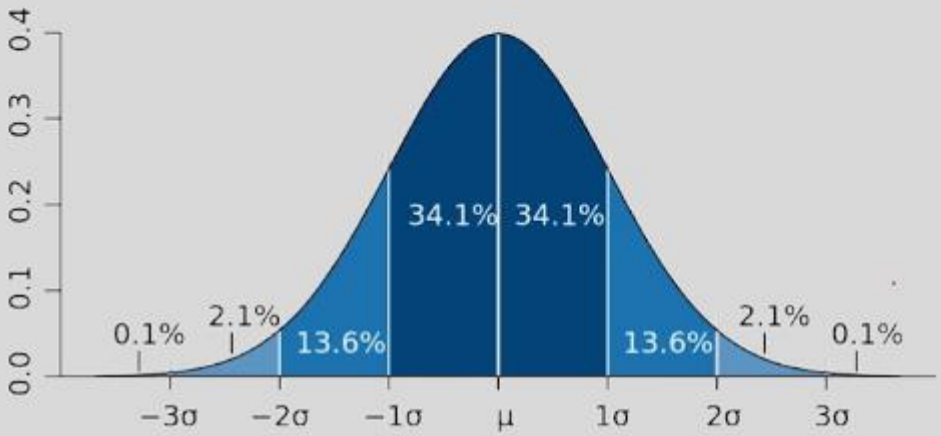
$$D = \frac{\sum(x_i - \bar{x})^2}{n} \qquad \delta = \sqrt{D} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$



ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Нормальное распределение

- Унимодально
- Симметрично



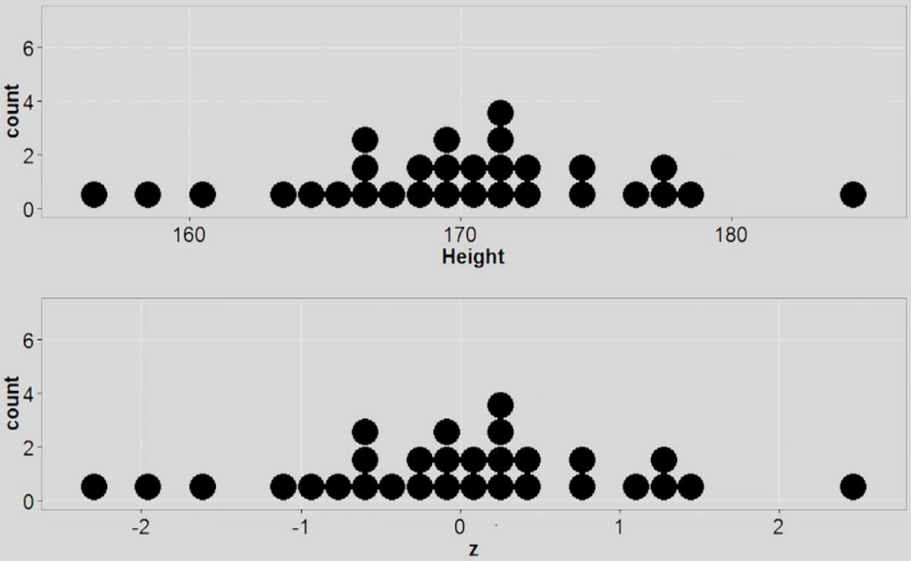
- Отклонения наблюдений от среднего подчиняются определенному вероятностному закону.

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Нормирование

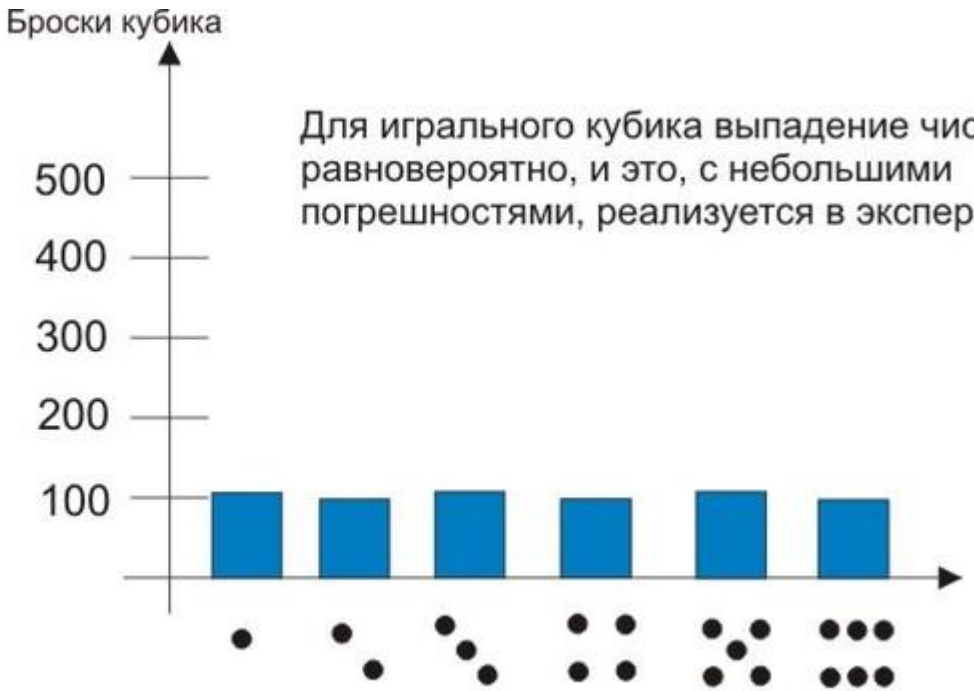
Стандартизация или *z-преобразование* – преобразование полученных данных в стандартную Z-шкалу (Z-scores) со средним $M_z = 0$ и $D_z = 1$

$$Z_i = \frac{x_i - \bar{x}}{\delta_x}$$



ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Равномерное распределение

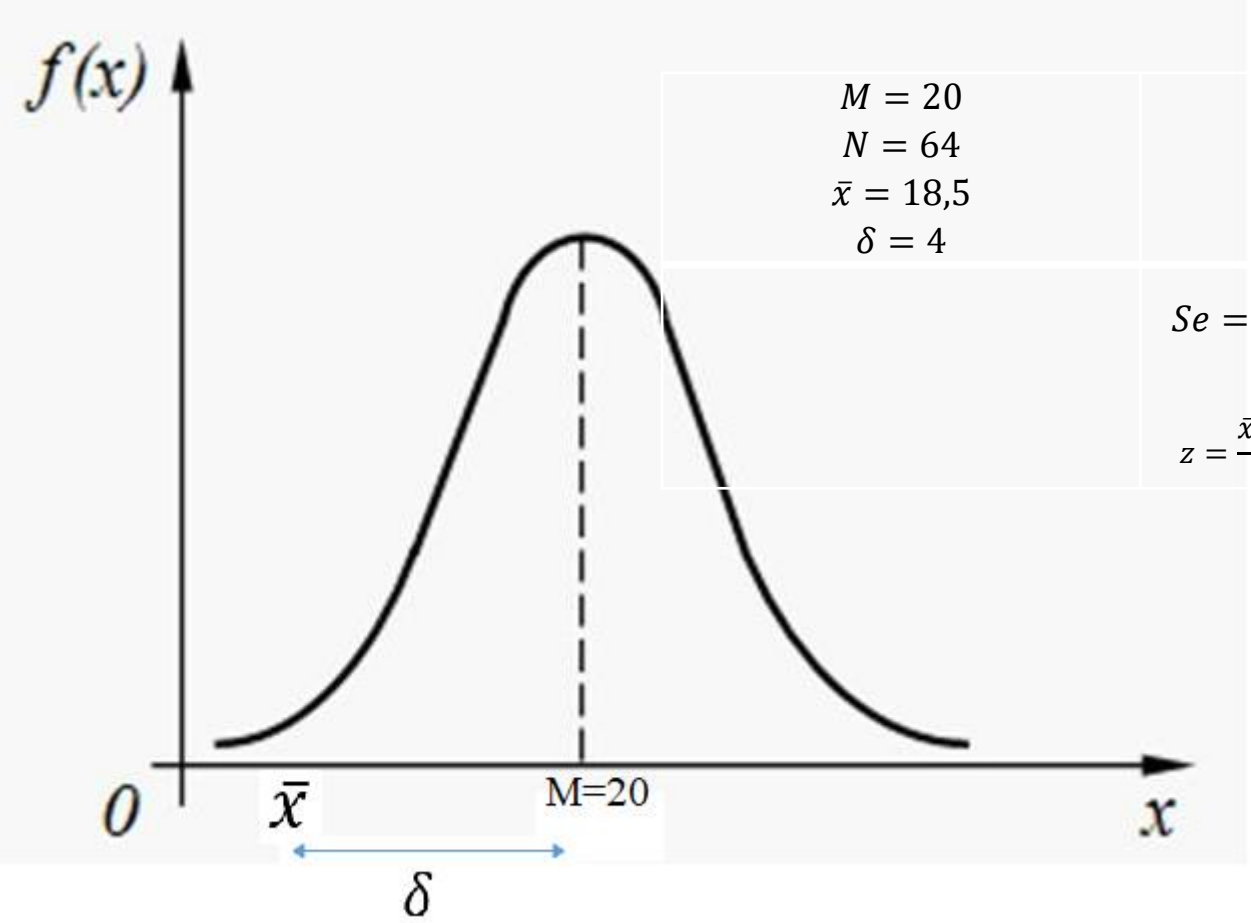


Для игрального кубика выпадение чисел равновероятно, и это, с небольшими погрешностями, реализуется в эксперименте.

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Статистическая проверка гипотез



$M = 20$

$N = 64$

$\bar{x} = 18{,}5$

$\delta = 4$

H_0: $M_{new} = 20$

H_1: $M_{new} \neq 20$

$$Se = \frac{\delta}{\sqrt{n}} = \frac{4}{\sqrt{64}} = \frac{4}{8} = 0{,}5$$

$$z = \frac{\bar{x} - M}{Se} = \frac{18{,}5 - 20}{0{,}5} = -3$$

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Статистическая проверка гипотез

https://gallery.shinyapps.io/dist_calc/

Distribution Calculator

Distribution:

Normal

Mean

-50 0 50

-50 -40 -30 -20 -10 0 10 20 30 40 50

Standard deviation

1 30

0.1 3.1 6.1 9.1 12.1 15.1 18.1 21.1 24.1 27.1 30

Model:

P(X < a or X > b)

Find Area:

Both Tails

a

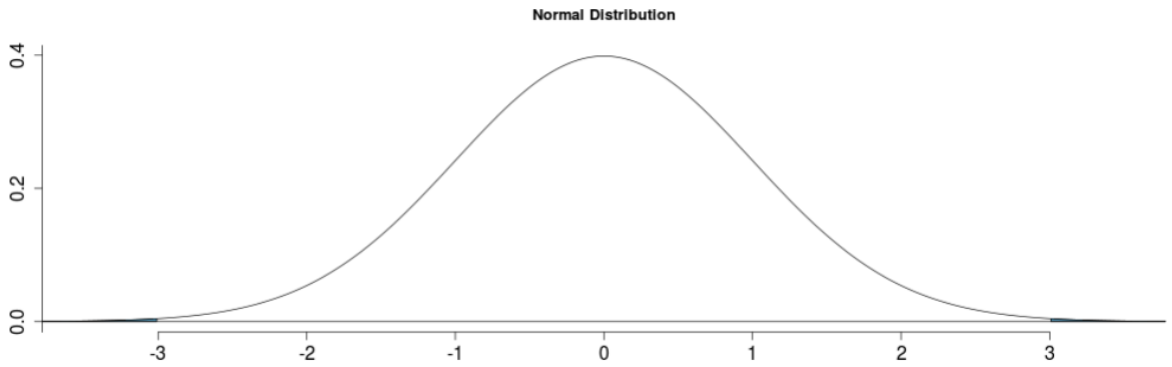
-4 -3 4

-4 -3.2 -2.4 -1.6 -0.8 0 0.8 1.6 2.4 3.2 4

b

-4 3 4

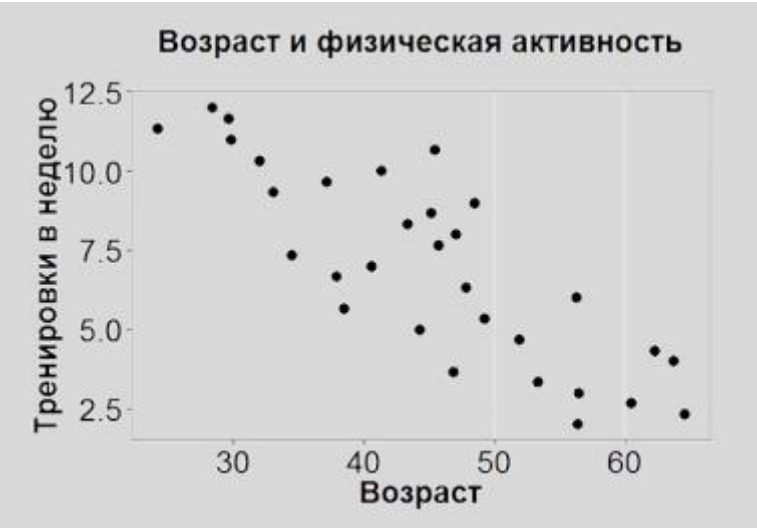
-4 -3.2 -2.4 -1.6 -0.8 0 0.8 1.6 2.4 3.2 4

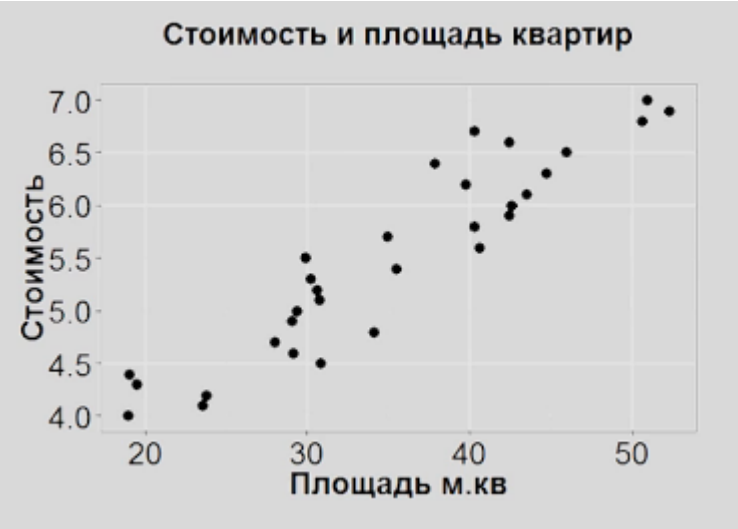


P(X < -3 or X > 3) = 0.0027

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Коэффициент корреляции

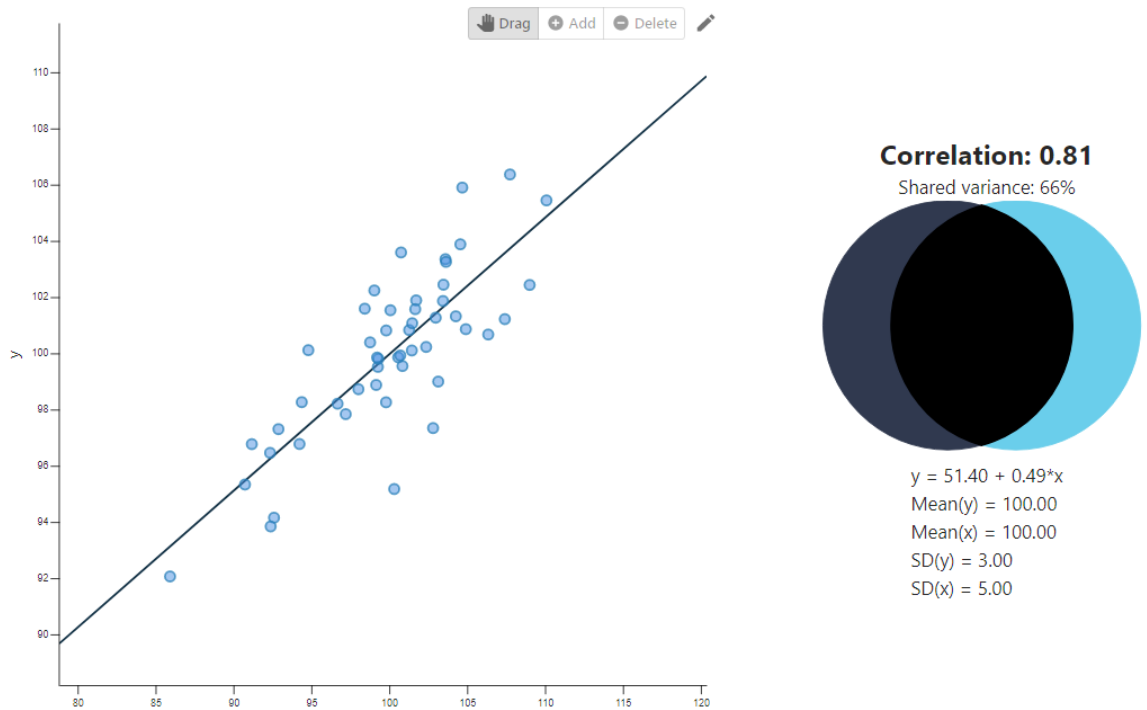




ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Коэффициент корреляции

https://rpsychologist.com/correlation/

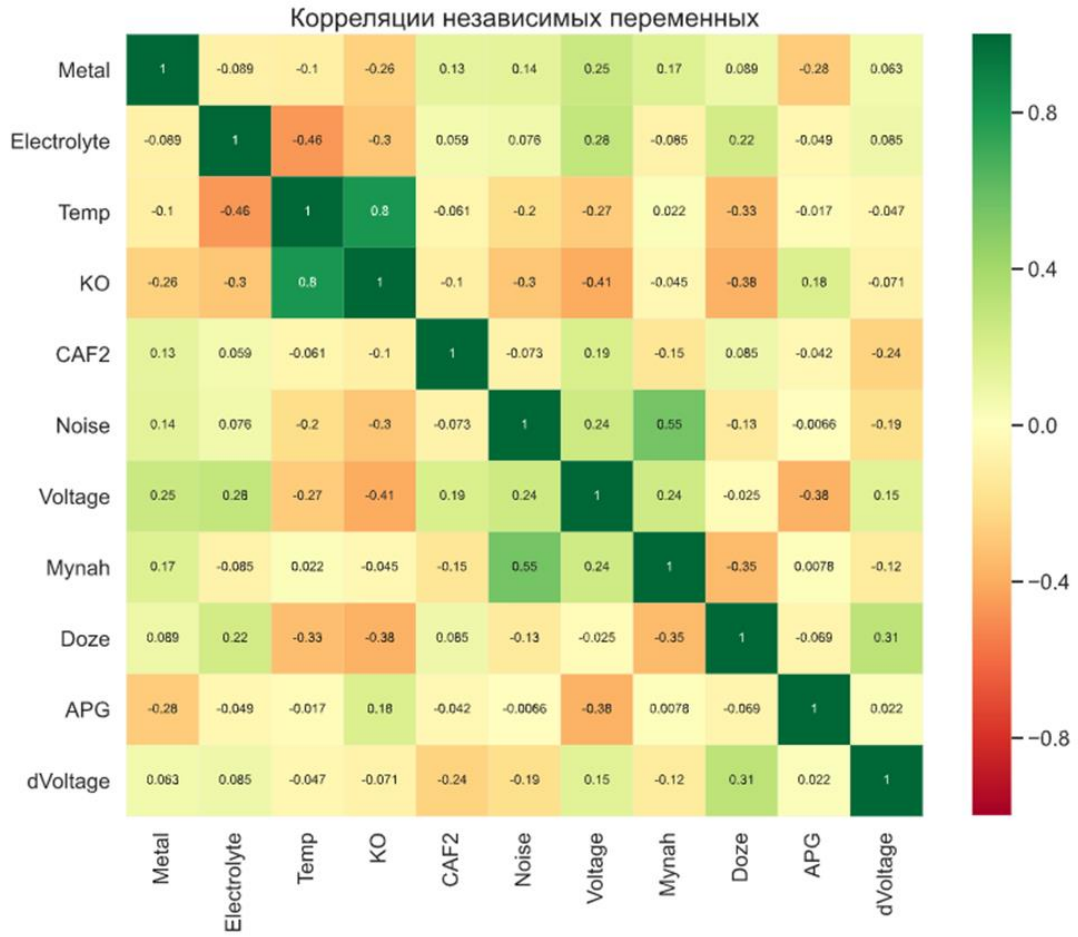


$$r_{xy} = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Коэффициент корреляции

Тепловая карта



ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

edu.bmstu.ru
+7 (495) 120-30-75
E-mail: edu@bmstu.ru
Москва, ул. 2-я Бауманская,
дом 5, стр. 1