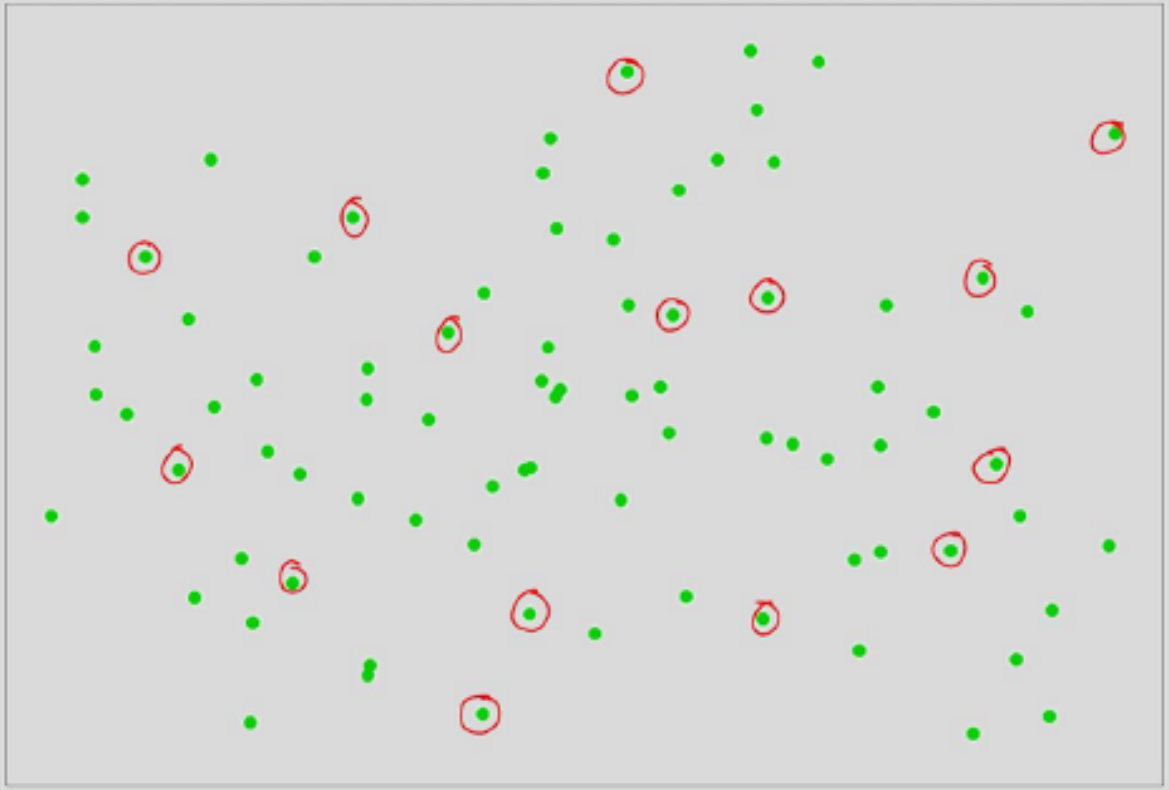
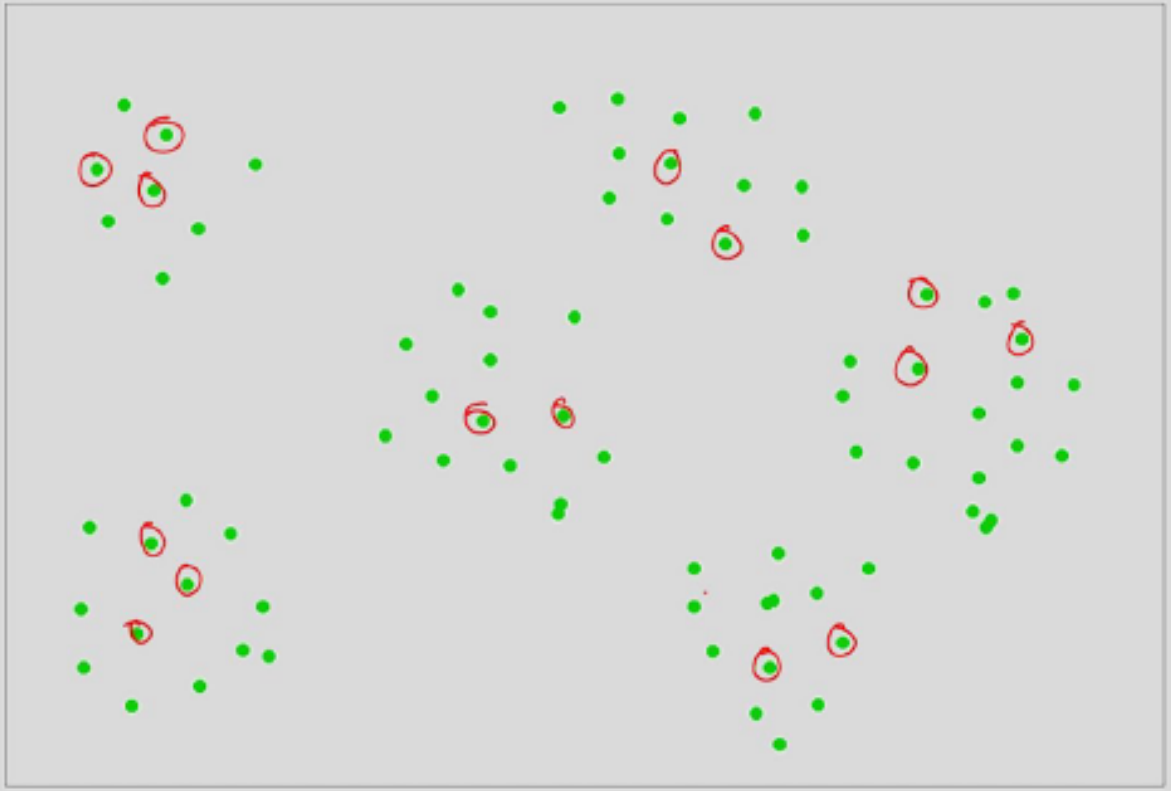
Выборка

- Простая случайная выборка (simple random sample)



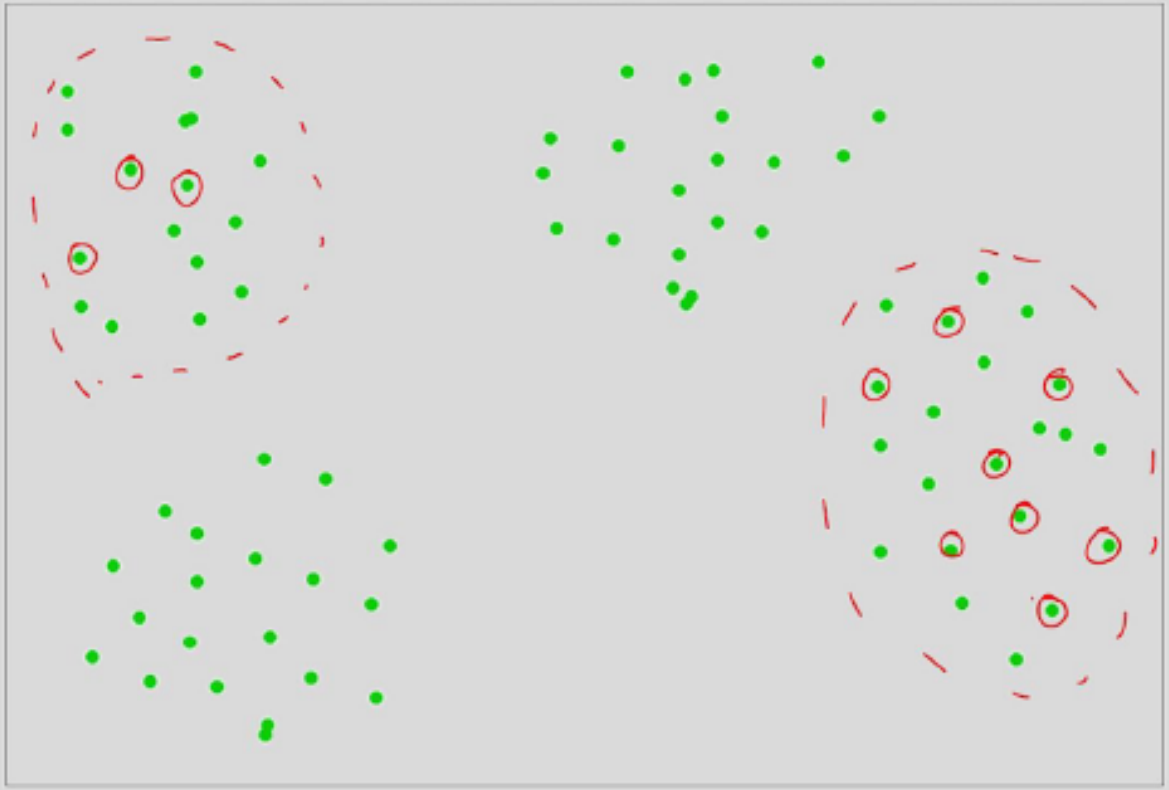
Выборка

- Стратифицированная выборка (stratified sample)



Выборка

- Групповая выборка (cluster sample)



Типы переменных

Рост

Количественные

- непрерывные [160; 190]
- дискретные

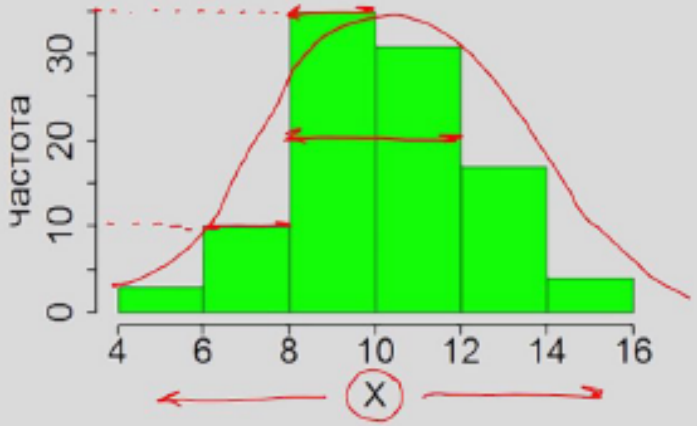
1 2 3 4

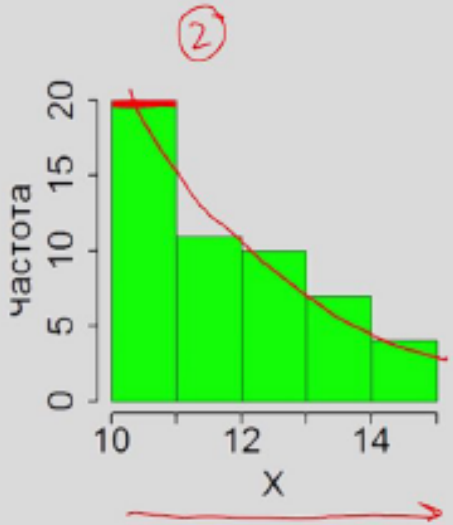
Номинативные

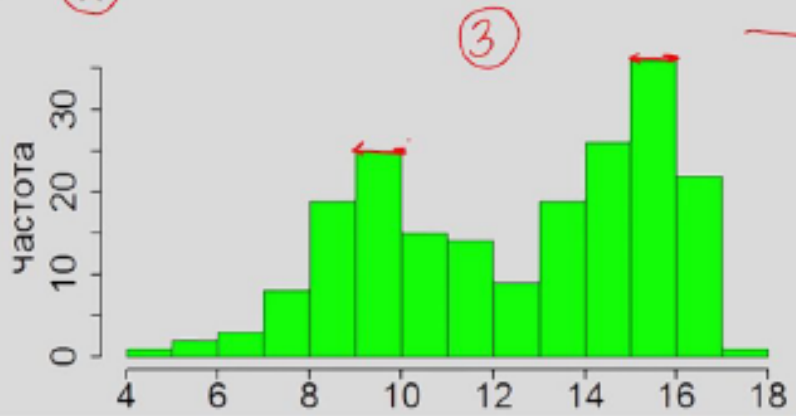
1 – ж

2 – м.

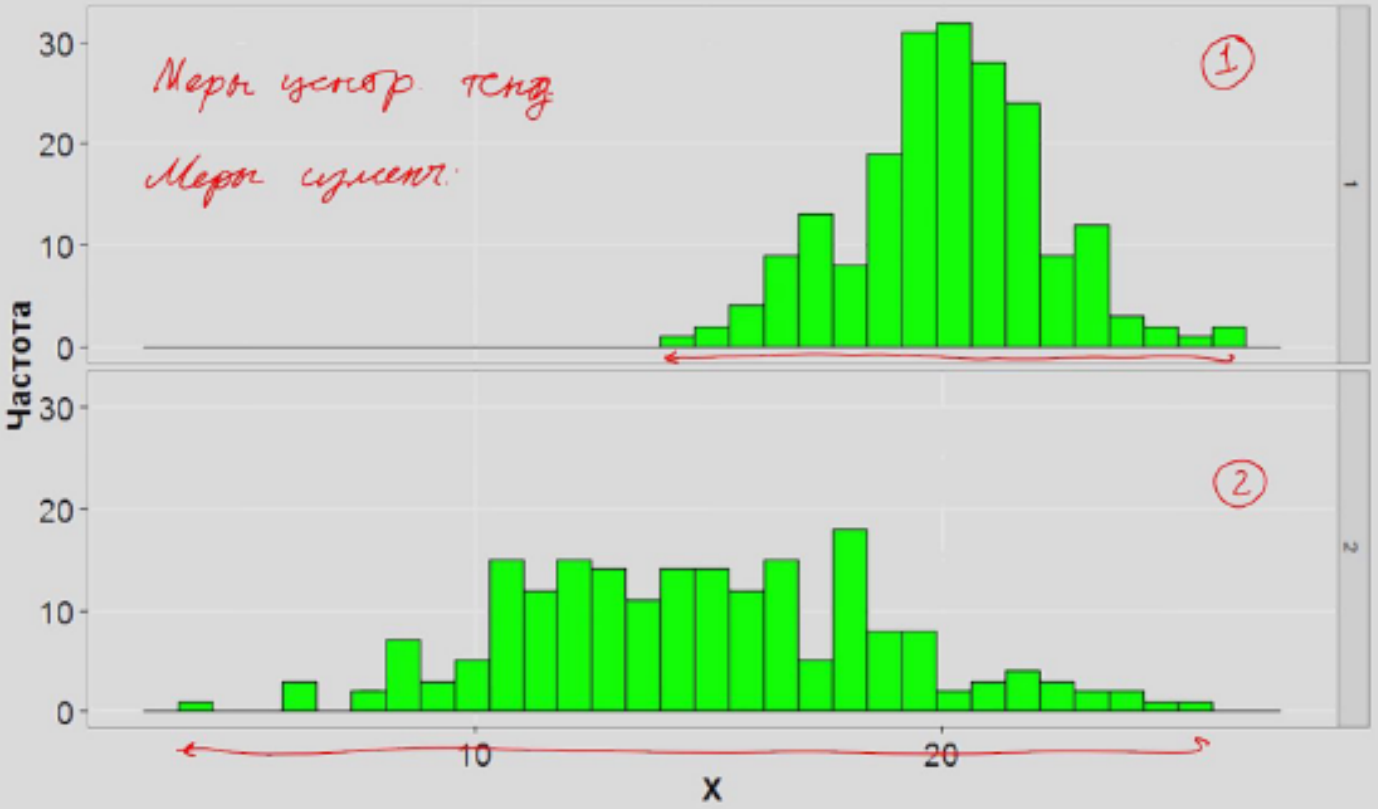
Гистограмма частот







Описательные статистики



Меры центральной тенденции

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

N = 30

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

Меры центральной тенденции

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

$N = 9$

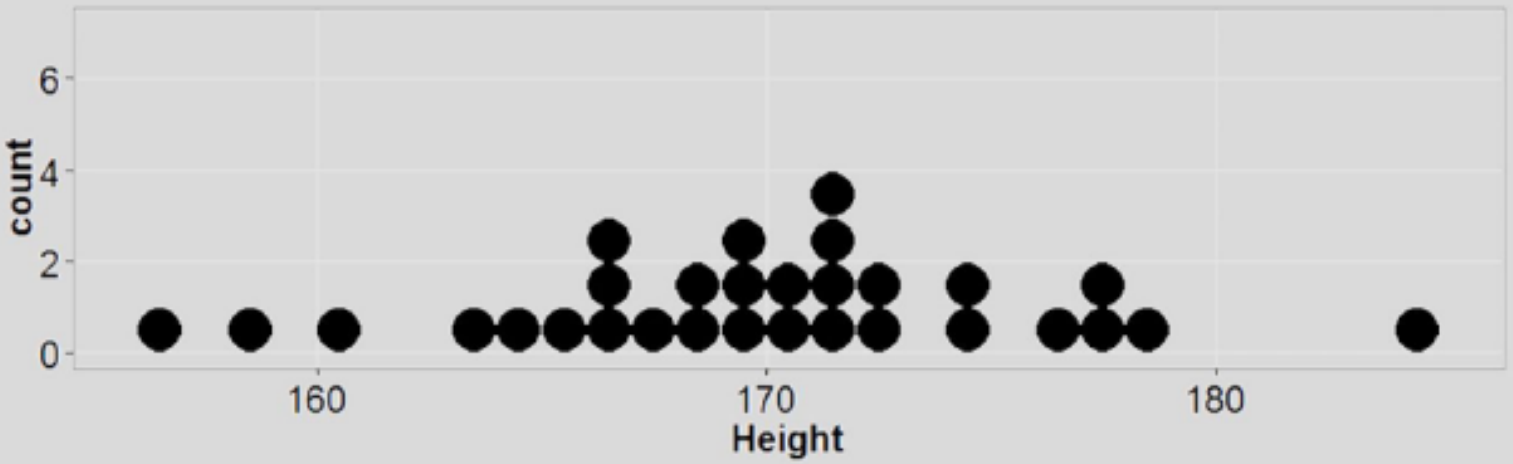
157 159 161 164 165 166 167 167 167

Медиана

$N = 30$

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

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



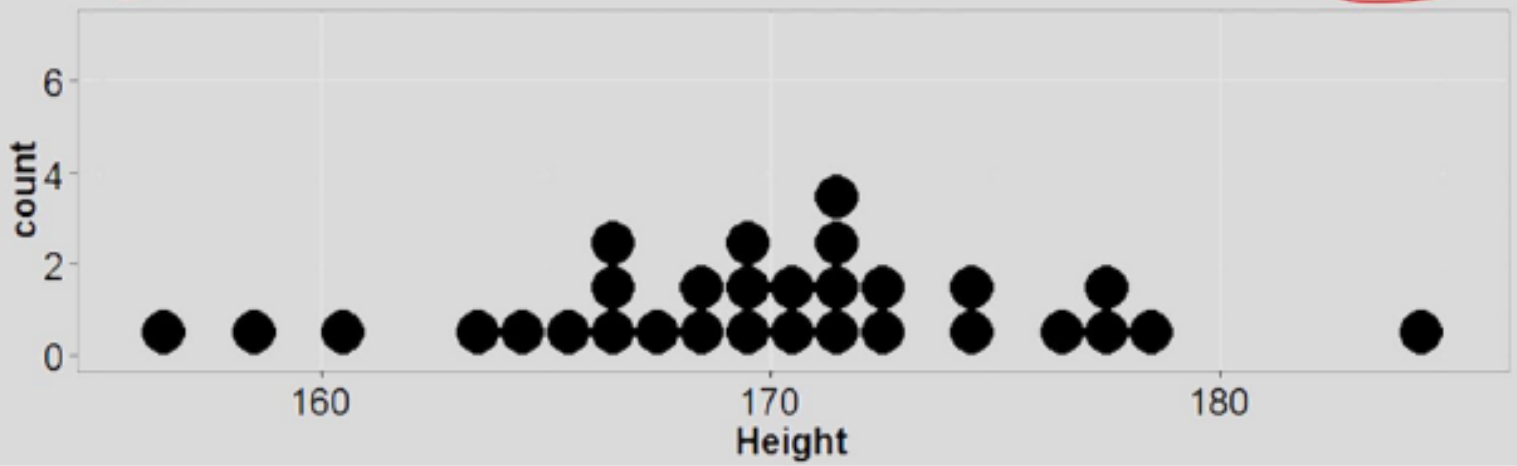
Меры центральной тенденции

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

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

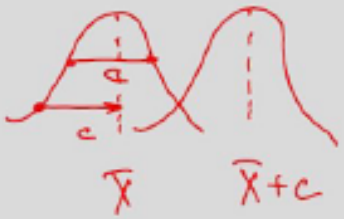
$\bar{X} = \frac{X_1 + \ldots + X_{30}}{30} = 170{,}4$

M



Свойства среднего

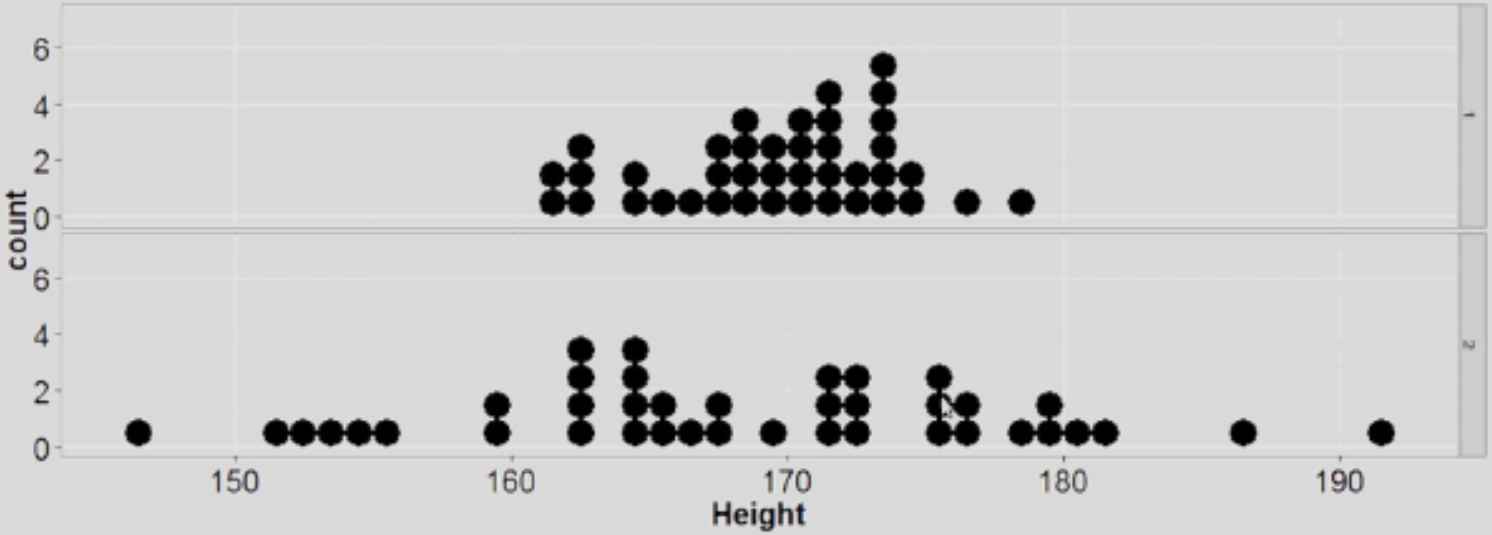
$$M_{x+c}=M_x+c$$



$$M_{x*c}=M_x*c$$

$$\Sigma(x_i-M_x)=0$$

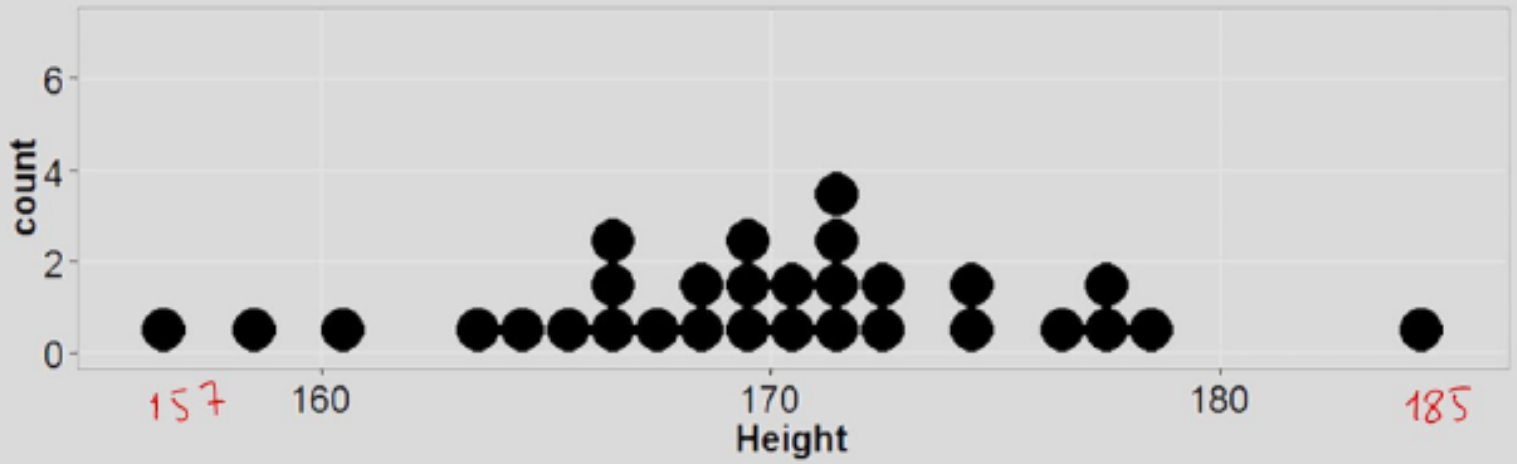
Меры изменчивости



Меры изменчивости

Размах (Range) - разность максимального и минимального значения.

$R = X_{max} - X_{min} = 185 - 157 = 28$

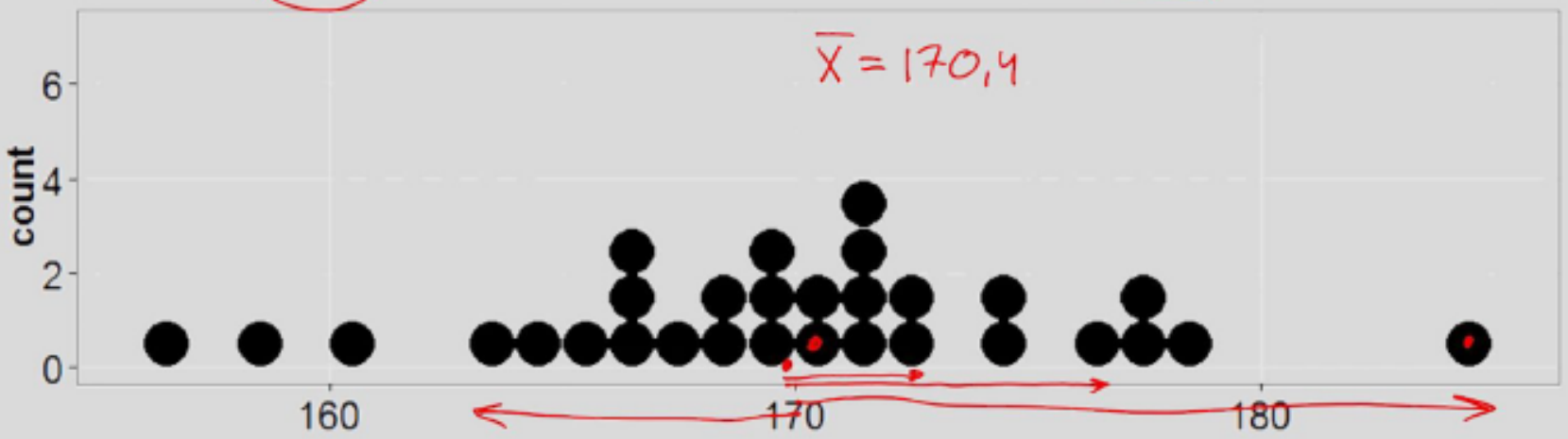


Меры изменчивости

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

ГС → Выборка $sd = 3{,}5$

$$D = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad \sqrt{D} = \sigma$$



$N = 7$

1 2 2 3 4 4 5

$Mx = 3$

$(1-3)^2 = 4$

$(2-3)^2 = 1$

$(2-3)^2 = 1$

$(3-3)^2 = 0$

$(4-3)^2 = 1$

$(4-3)^2 = 1$

$(5-3)^2 = 4$

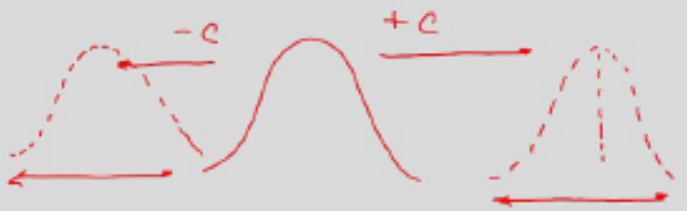
$$D = \frac{4+1+\ldots+1+4}{7-1} = 2$$

$$Sd = \sqrt{D} = 1.4$$

Свойства дисперсии

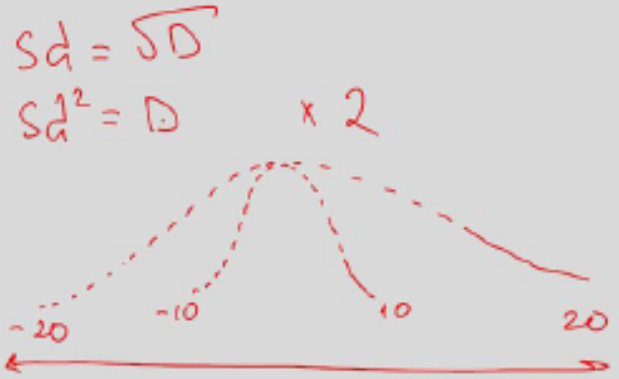
$$D_{x+c} = D_x$$

$$sd_{x+c} = sd_x$$



$$D_{x*c} = D_x * c^2$$

$$sd_{x*c} = sd_x * c$$



Квантили распределения

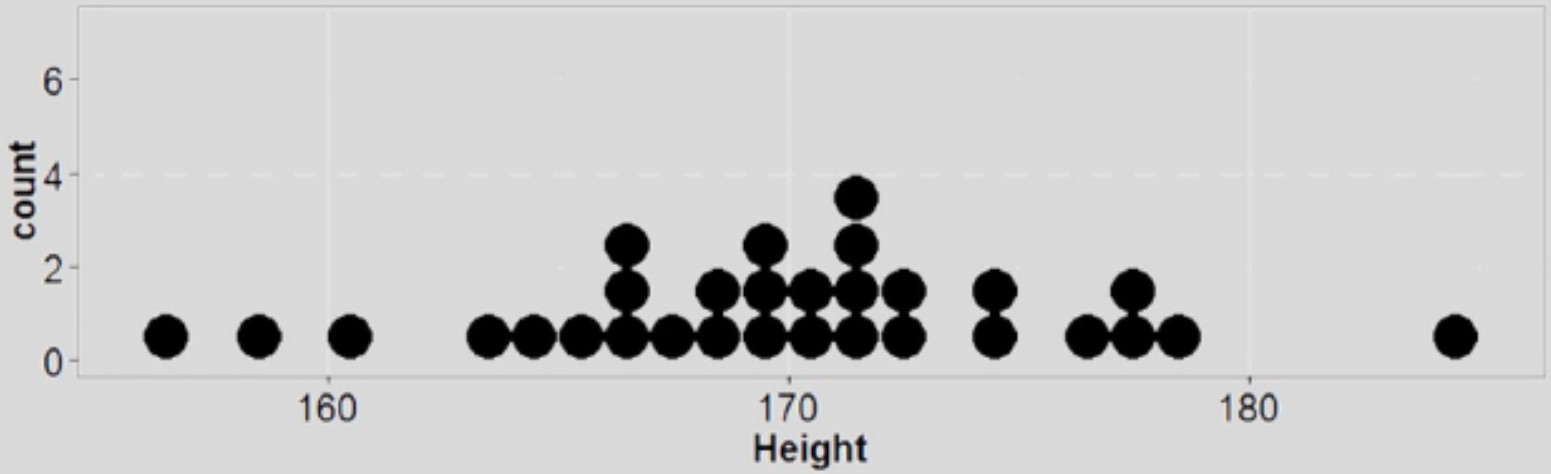
Квартили – три точки (значения признака), которые делит упорядоченное множество данных на четыре равные части.

1 квартиль

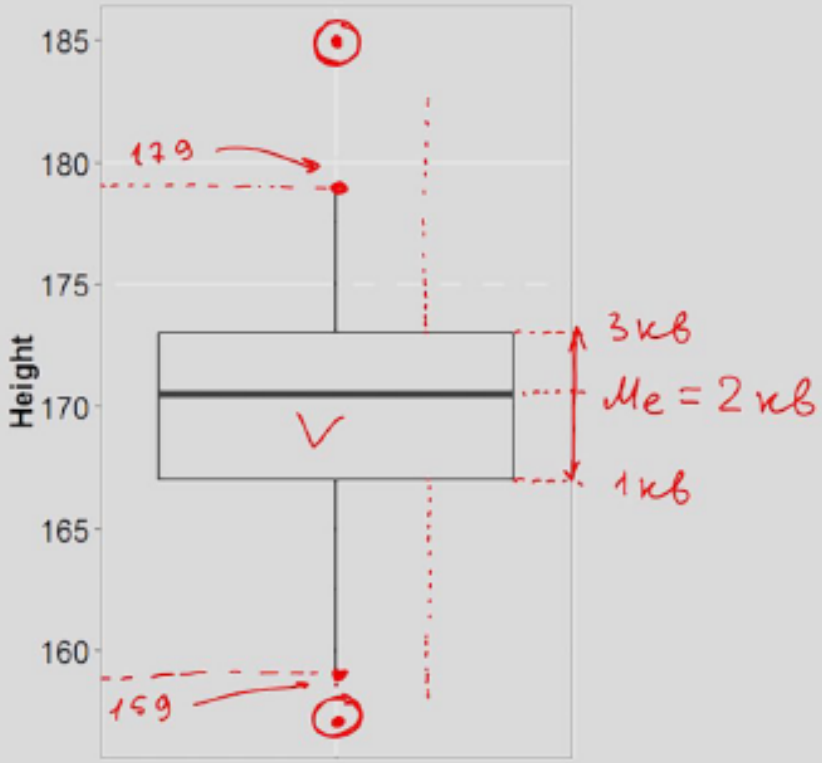
157 159 161 164 165 166 167 167 167 168 169 169 170 170 170
171 171 172 172 172 172 173 173 175 175 177 178 178 179 185

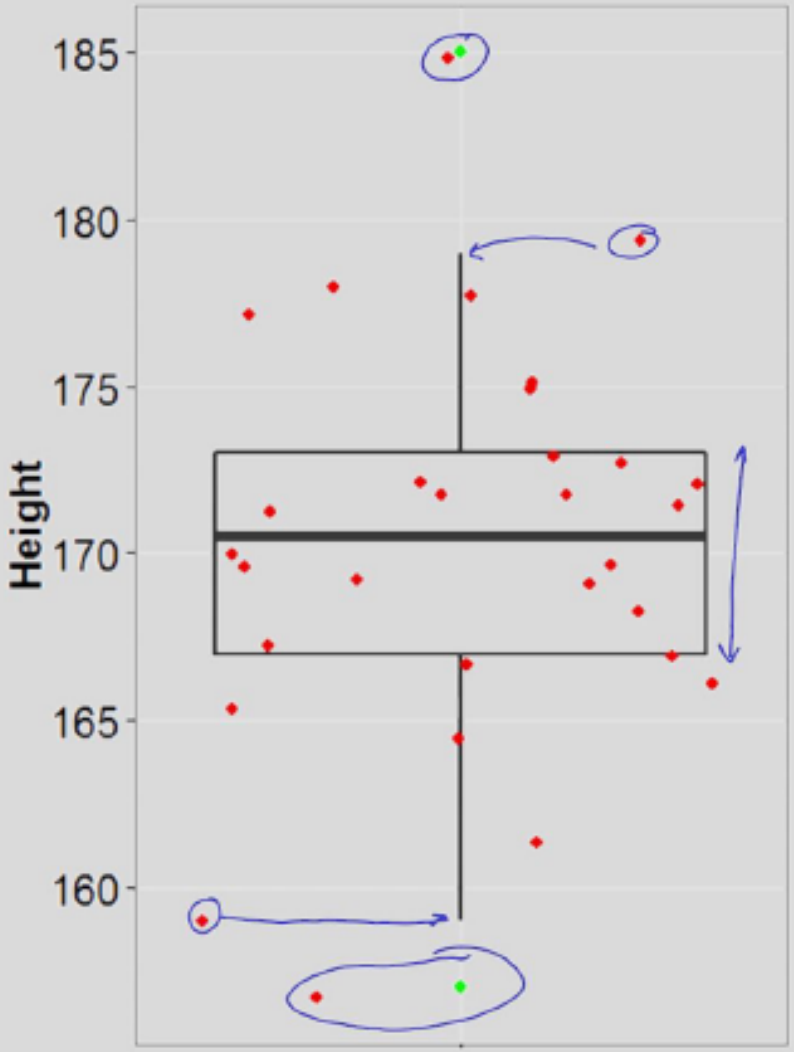
170,5

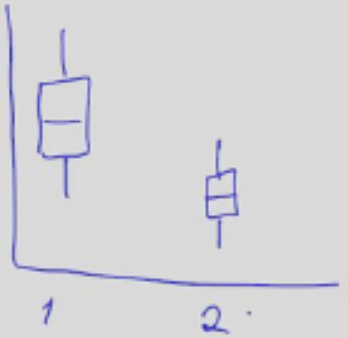
3 квартиль



Box plot

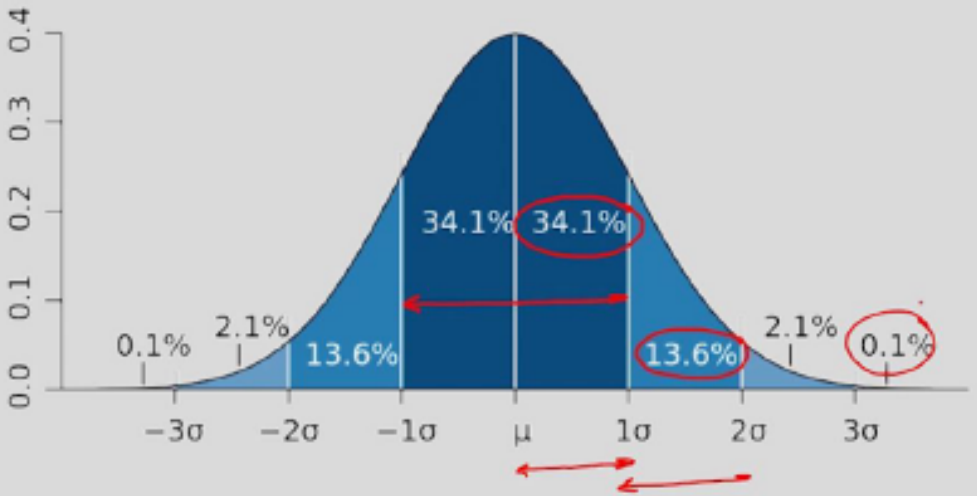



185
180
175
170
165
160
Height


1
2

Нормальное распределение

- Унимодально
- Симметрично



- Отклонения наблюдений от среднего подчиняются определенному вероятностному закону.

Стандартизация

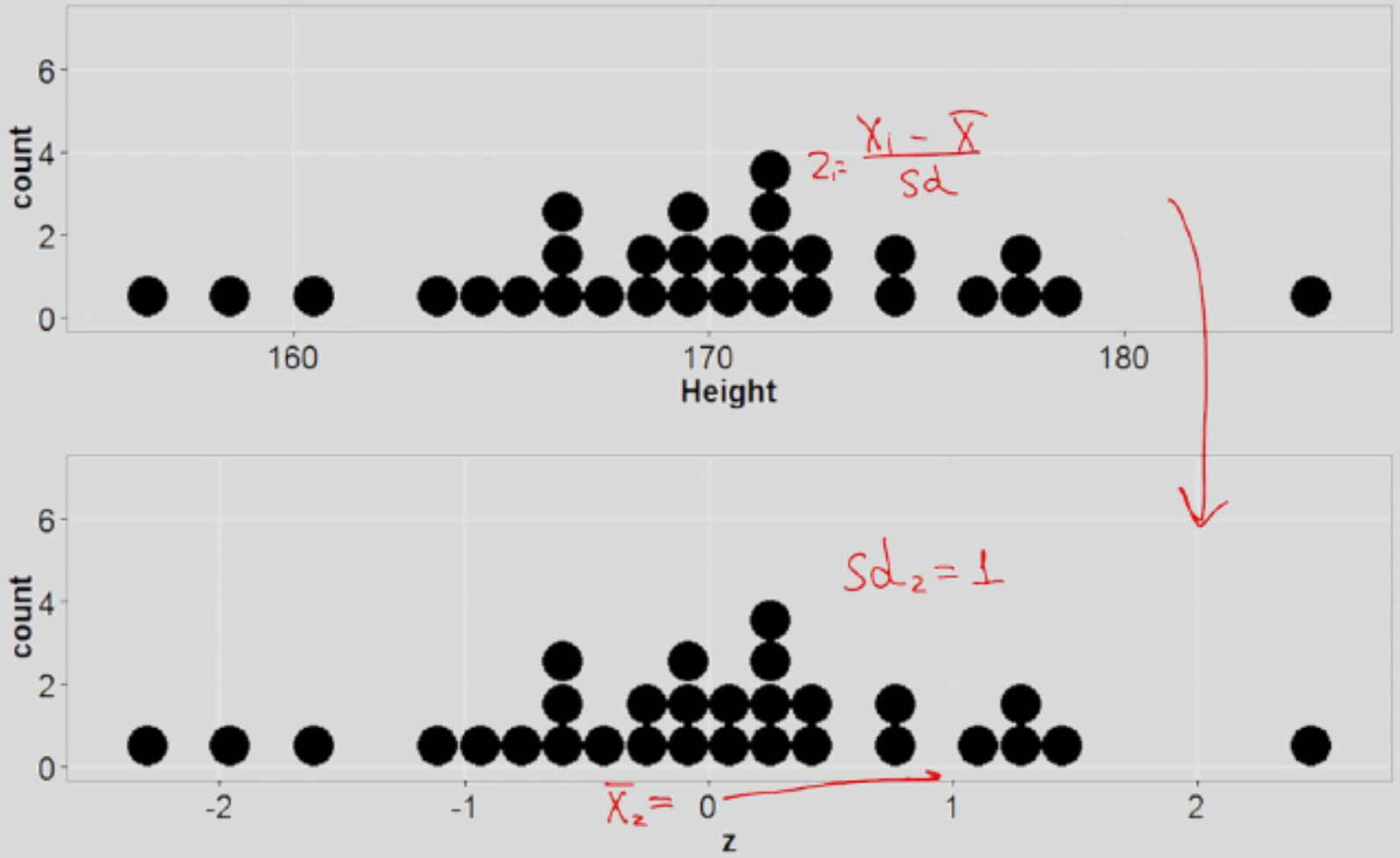
Стандартизация или *z-преобразование* – преобразование полученных данных в стандартную Z-шкалу (Z-scores) со средним $M_z = 0$ и $D_z = 1$

$$Z_i = \frac{X_i - \bar{X}}{\sigma_x}$$

$$\bar{X} - c = \bar{X} - \bar{X} = 0$$

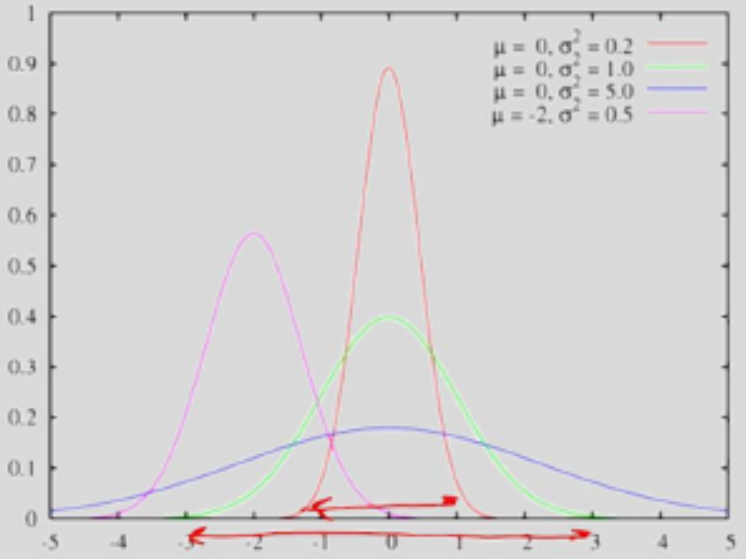
$$\cdot \frac{1}{\sigma_x} \qquad D_x \cdot \left(\frac{1}{\sigma_x}\right)^2 = D_x \cdot \frac{1}{D_x} = 1$$

Стандартизация



Правило «двух» и «трех» сигм

- $M_x \pm \sigma \approx 68\%$ наблюдений
- $M_x \pm 2\sigma \approx 95\%$ наблюдений
- $M_x \pm 3\sigma \approx 100\%$ наблюдений



z	0,00	0,01	0,02	0,03	0,04
0,0	0,5000	0,4960	0,4920	0,4880	0,4840
0,1	0,4602	0,4562	0,4522	0,4483	0,4404
0,2	0,4207	0,4168	0,4129	0,4090	0,4052
0,3	0,3821	0,3783	0,3745	0,3707	0,3669
0,4	0,3446	0,3409	0,3372	0,3336	0,3300
0,5	0,3085	0,3050	0,3015	0,2981	0,2946
0,6	0,2743	0,2709	0,2676	0,2643	0,2611
0,7	0,2420	0,2389	0,2358	0,2327	0,2296
0,8	0,2119	0,2090	0,2061	0,2033	0,2005
0,9	0,1841	0,1814	0,1788	0,1762	0,1736
1,0	0,1587	0,1562	0,1539	0,1515	0,1492

$\bar{X} = 150$
$sd = 8$
154

$$Z = \frac{154 - 150}{8} = 0{,}5$$

Central Limit Theorem for Means

Parent distribution (population):

- (•) Normal
- ( ) Uniform
- ( ) Right skewed
- ( ) Left skewed

Mean

Standard deviation

Sample size:

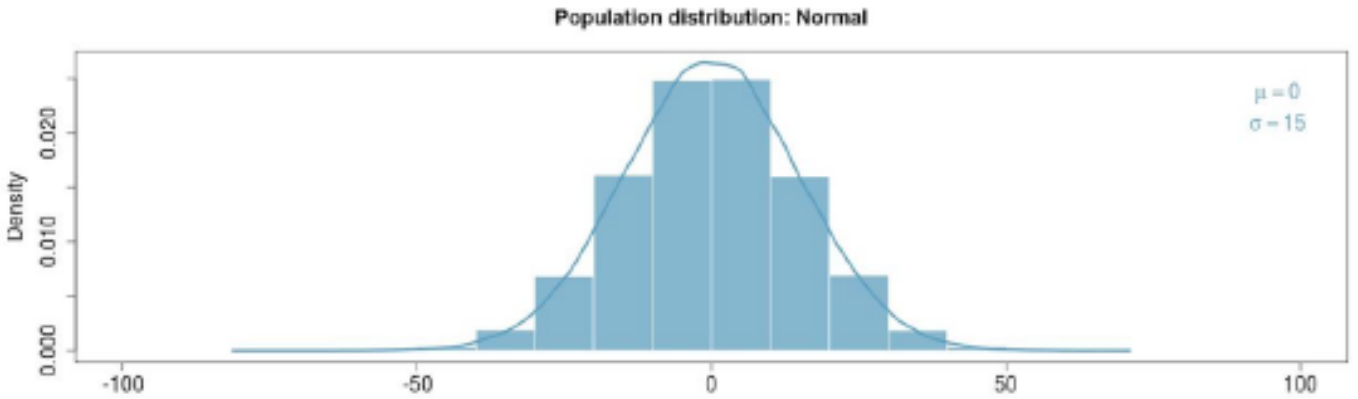
Number of samples:

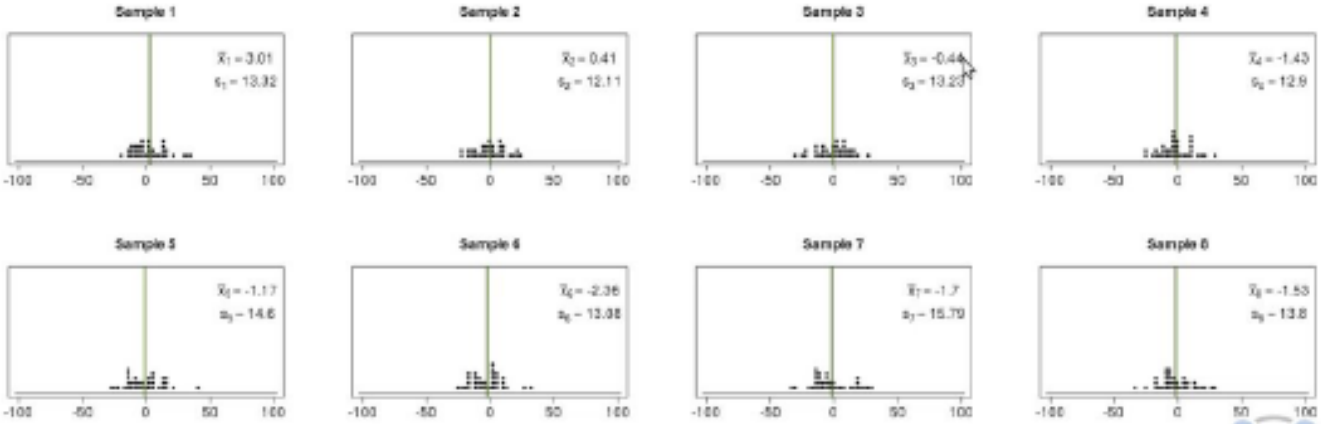
Rate this app!

View code

Check out other apps

Want to learn more for free?





... continuing to Sample 501.

Distribution Calculator | Central Limit Theorem ...

spark.rstudio.com/minebocek/CLT_mean/

Google

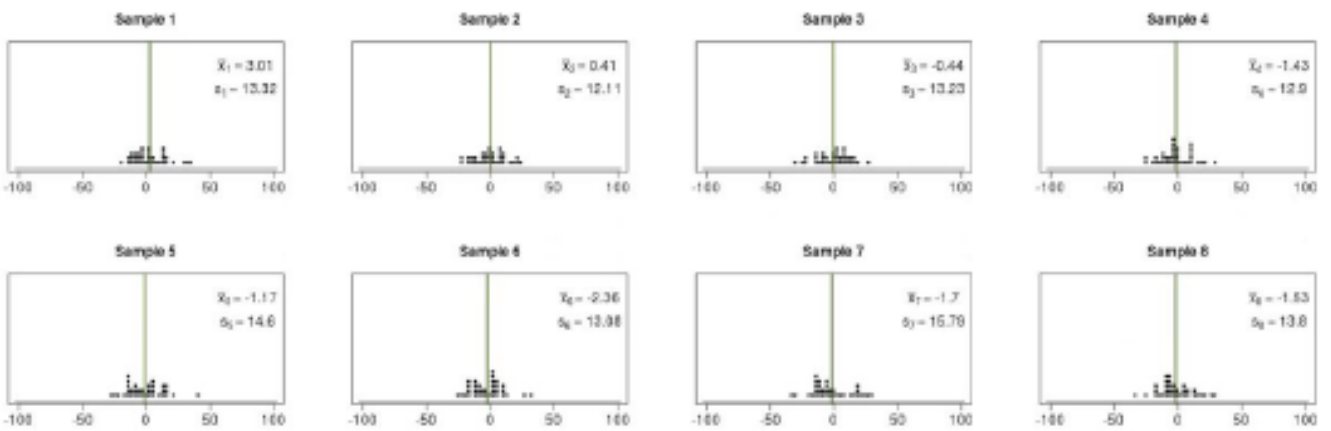
10 501 1,000

Rate this app!

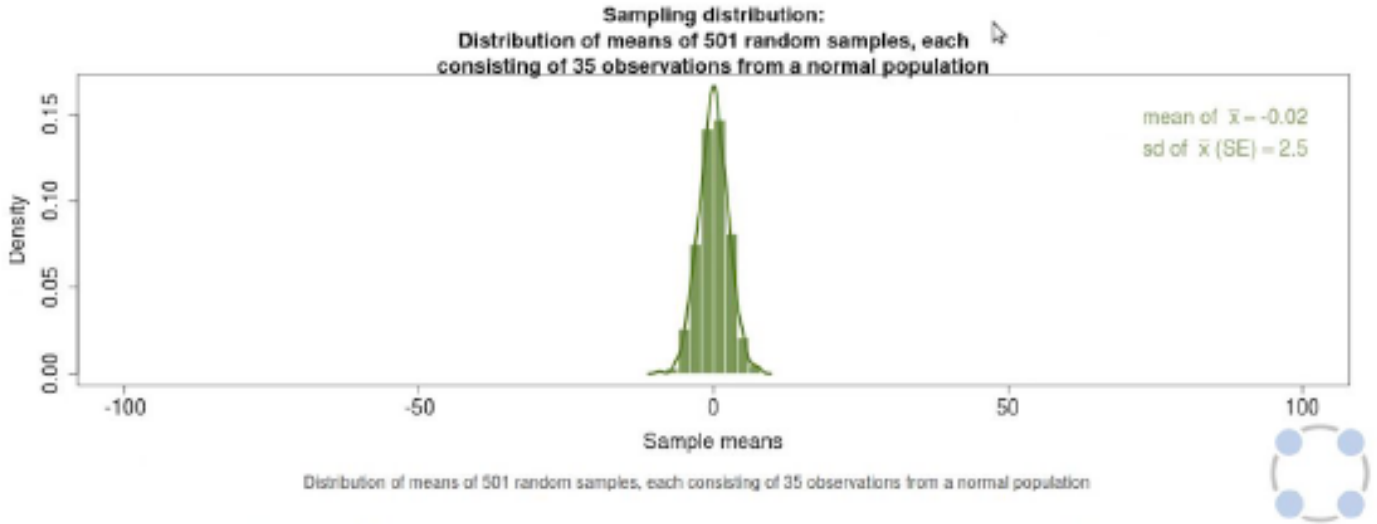
View code

Check out other apps

Want to learn more for free?

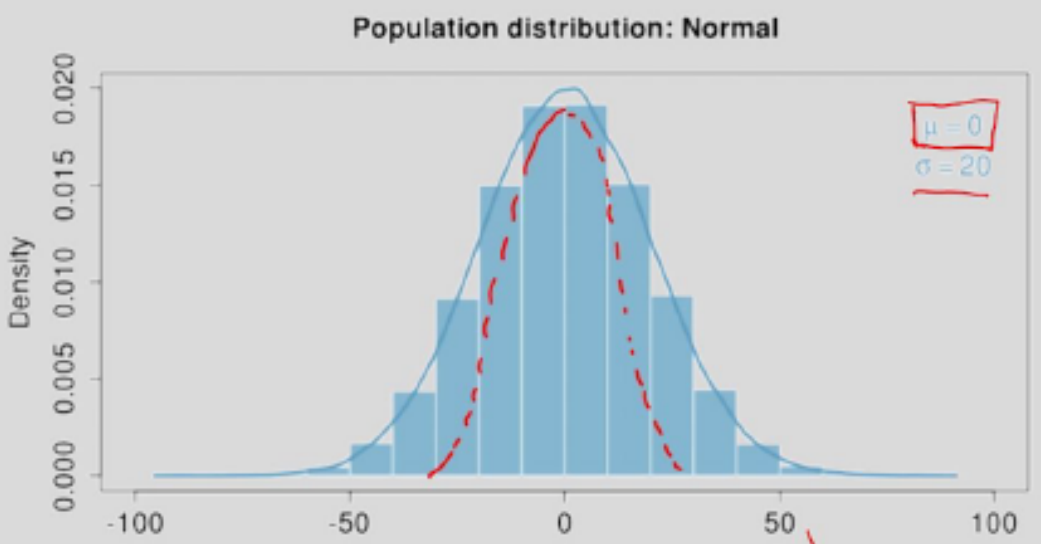


... continuing to Sample 501.



Distribution of means of 501 random samples, each consisting of 35 observations from a normal population

According to the Central Limit Theorem (CLT), the distribution of sample means (the sampling distribution) should be nearly normal. The mean of the sampling distribution



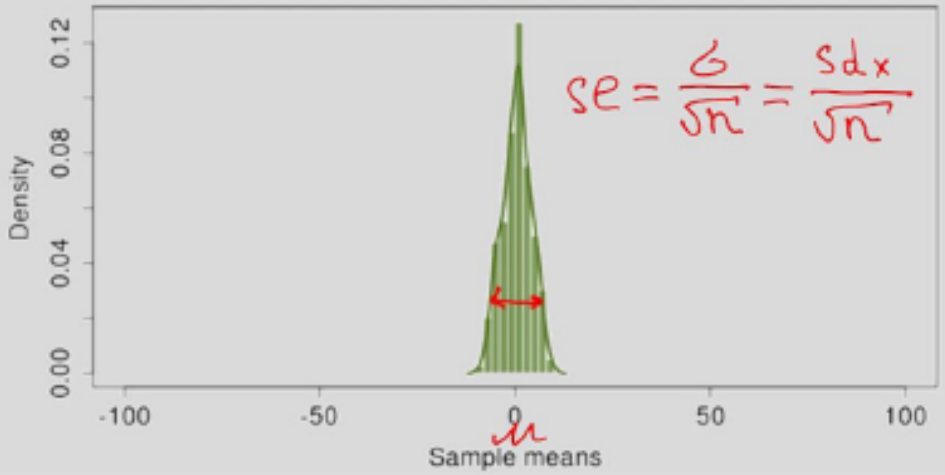
$n > 30$

$\bar{X}_1$

$\bar{X}_2$

$\vdots$

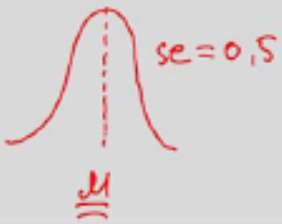
$\bar{X}_n$



$se = \frac{\sigma}{\sqrt{n}} = \frac{sd_x}{\sqrt{n}}$

$n = 100 \quad sd = 5 \quad \bar{X} = 3$

$se = \frac{5}{\sqrt{100}} = 0{,}5$



Построение доверительных интервалов

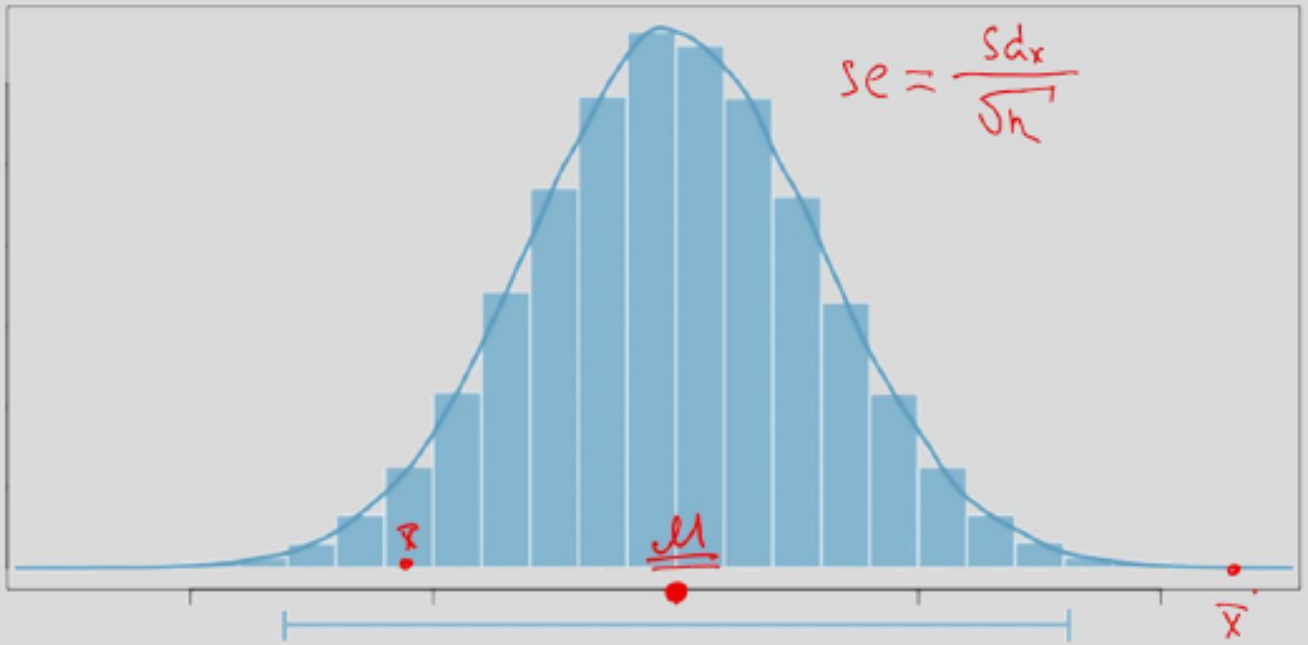
Уровень экспрессии некоторого гена измерялся в эксперименте. Ниже представлены результаты 64 наблюдений.

[18:30]

102 91 99 100 103 98 99 101 106 88 103 97 103 101
101 91 104 105 105 100 101 91 99 98 107 102 100 97
98 104 100 98 102 99 95 103 104 97 99 102 98 107 101
93 98 101 93 91 107 102 96 93 100 105 103 107 99 102
106 102 94 104 103 102

$\bar{X}=100$ $sd=4$

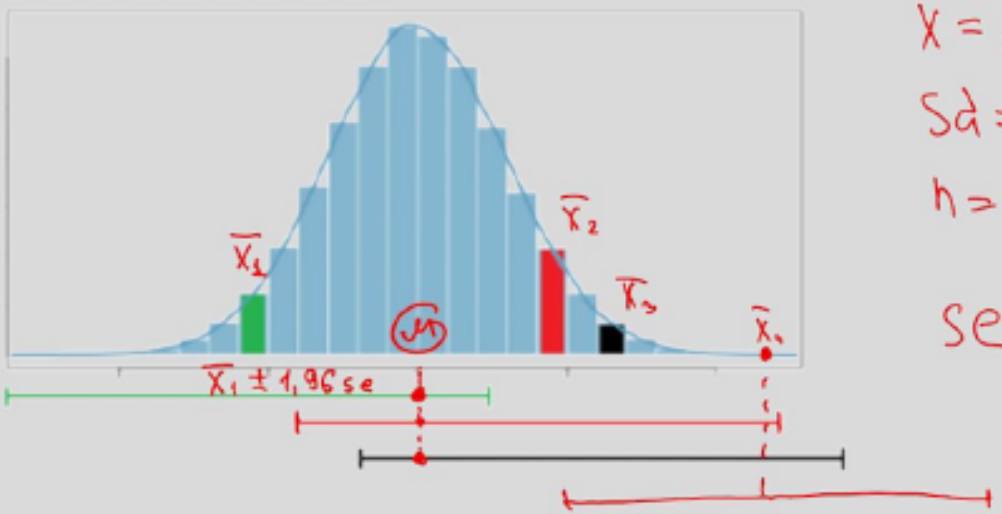
Построение доверительных интервалов



$$\mu \pm 1{,}96\sigma$$

$$\pm 1{,}96\,se$$

Построение доверительных интервалов



$\bar{X} = 100$

$Sd = 4$

$n = 64$

$$se = \frac{sd_x}{\sqrt{n}} = \frac{4}{\sqrt{64}} = \frac{4}{8} = 0{,}5$$

$\bar{X} - 1{,}96\,se$ ⟵ $\bar{X}$ ⟶ $\bar{X} + 1{,}96\,se$

$100 - 1{,}96 \cdot 0{,}5$

$100 - 0{,}98$

$100 + 1{,}96 \cdot 0{,}5$

$100 + 0{,}98$

$99{,}02$ ⟵ $\bar{X}$ ⟶ $100{,}98$

Идея статистического вывода

$M = 20$

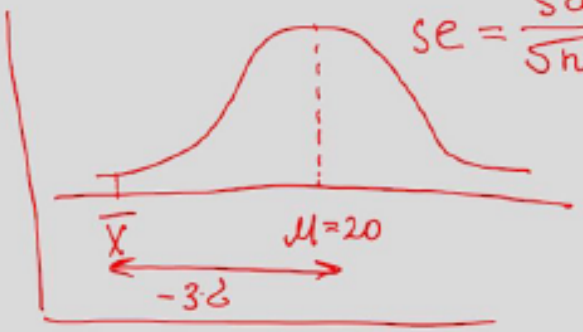
$N = 64$

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

$sd = 4$

$H_0 \quad \mu_{нп} = 20$

$H_1 \quad \mu_{нп} \neq 20$



$$se = \frac{sd}{\sqrt{n}} = \frac{4}{\sqrt{64}} = 0{,}5$$

$$z = \frac{\bar{X} - \mu}{se} = \frac{18{,}5 - 20}{0{,}5} = -3$$

Distribution Calculator

spark.rstudio.com/minebocek/dist_calc/

Distribution Calculator

Distribution:

Normal

Mean

−50 0 50

Standard deviation

1 30

Model:

P(X < a or X > b)

Find Area:

Both Tails

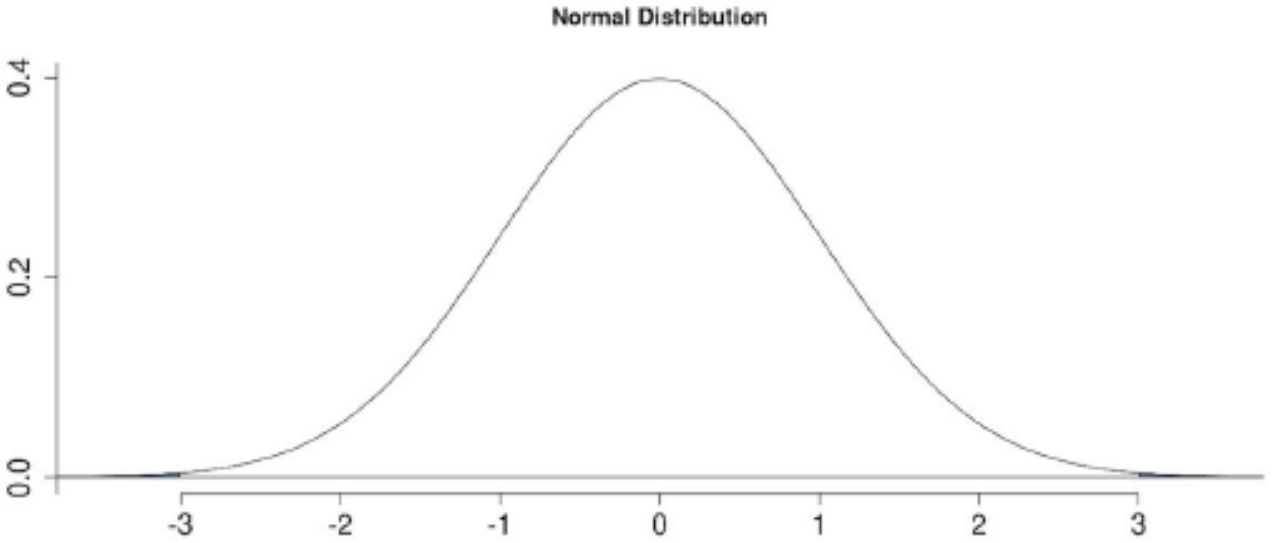
a

−4 −3 4

b

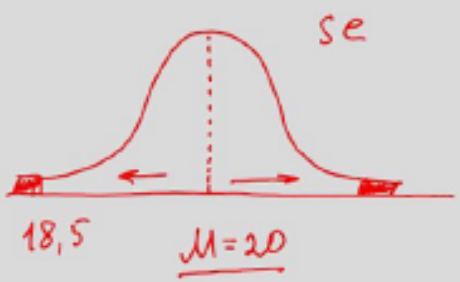
−4 3 4

Rate this app!



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

Идея статистического вывода



$p = 0{,}003$

H_0: $M_{нп} = 20$ $p > 0{,}05$

H_1: $M_{нп} \neq 20$ $p < 0{,}05$!!!