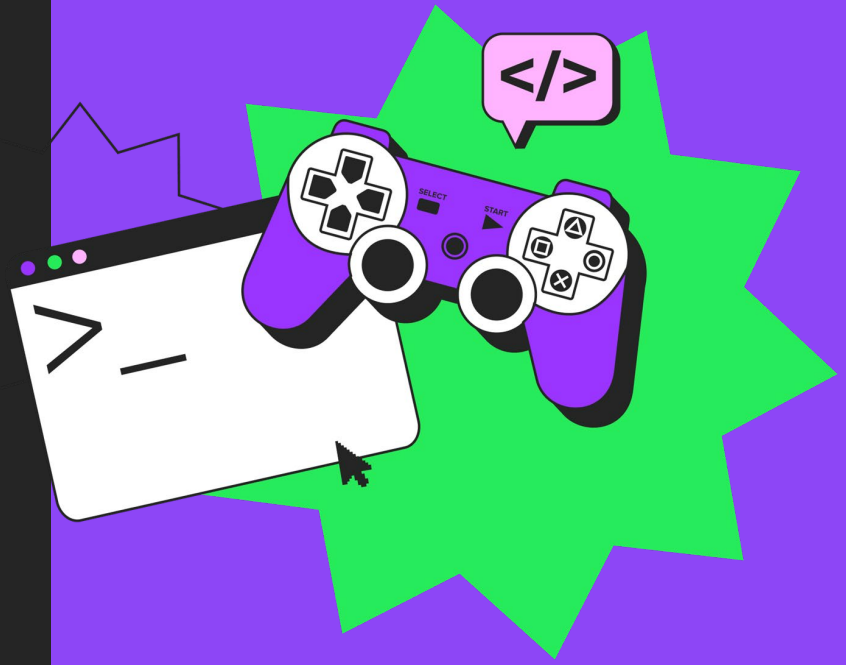
GeekBrains

Сравнение долей. Построение доверительных интервалов.

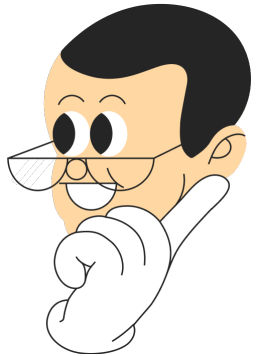


План курса

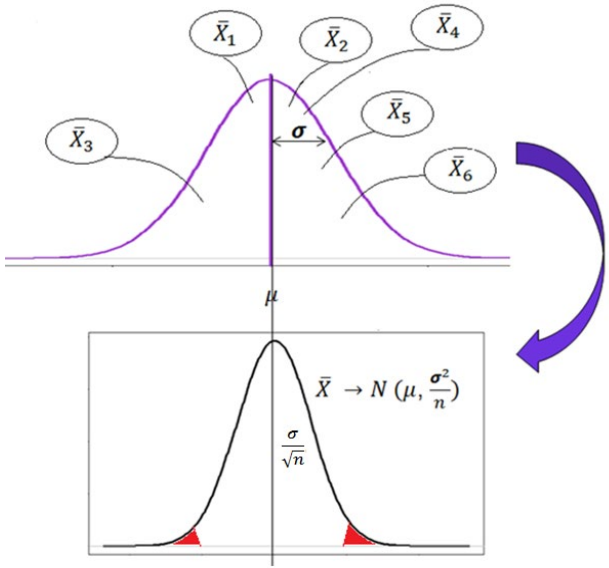
1. Расчет вероятностей
2. Дискретные распределения вероятностей
3. Разведочный анализ
4. Непрерывная случайная величина
5. Тестирование гипотез. Параметрические тесты
6. Сравнение долей. Построение доверительных интервалов
7. Будущий урок
8. Будущий урок
9. Будущий урок
10. Будущий урок
11. Будущий урок
12. Будущий урок
13. Прошедший урок
14. Сегодняшний урок
15. Будущий урок
16. Будущий урок

Что будет на уроке сегодня

- 📌 Доверительный интервал для средних арифметических
- 📌 Интервальная оценка для разности средних арифметических
- 📌 Доверительный интервал для доли
- 📌 Маленькие объемы выборок
- 📌 Сравнение долей
- 📌 Интервал для разности долей



Какую задачу решает доверительный интервал?

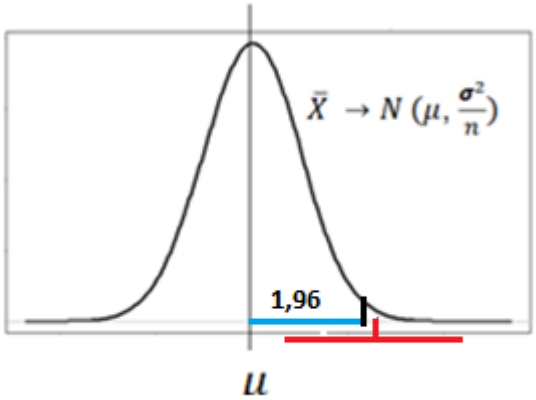


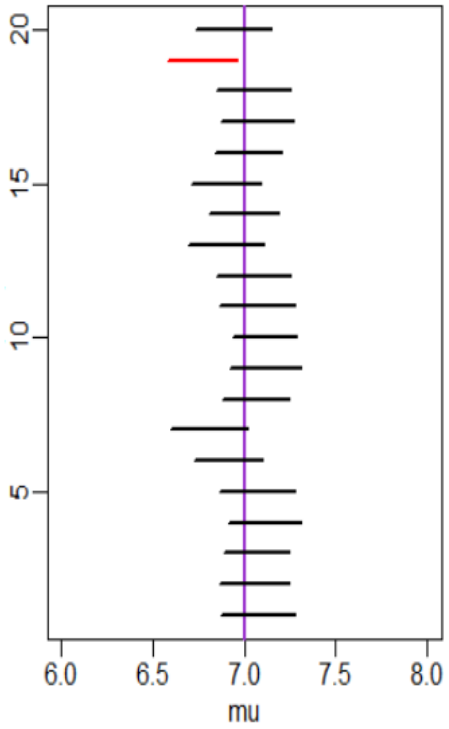
Доверительный интервал для средних арифметических

(1-α) = 0.95 , т.е. α =0.05

$\bar{x} \pm z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$, если σ генеральной совокупности известна,

$\bar{x} \pm t_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$, если σ генеральной совокупности неизвестна, тогда вычисляем σ по выборке, используя формулу для несмещенного стандартного отклонения, или функция в Python std (x, ddf =1).





Доверительный интервал. Сигма совокупности известна

Известно, что генеральная совокупность распределена нормально со средним квадратичным 5. Найти доверительный интервал для оценки среднего арифметического с надежностью 0,95, если выборочная средняя M =24.15, а объем выборки 100.

$$\overline{X} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

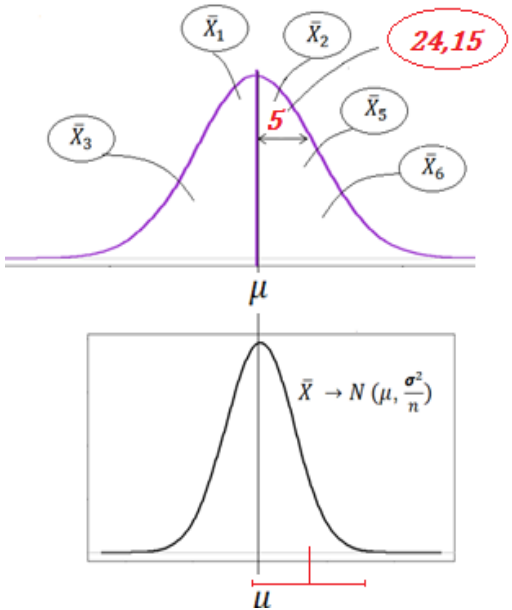


Таблица Z – значения (левая часть распределения)

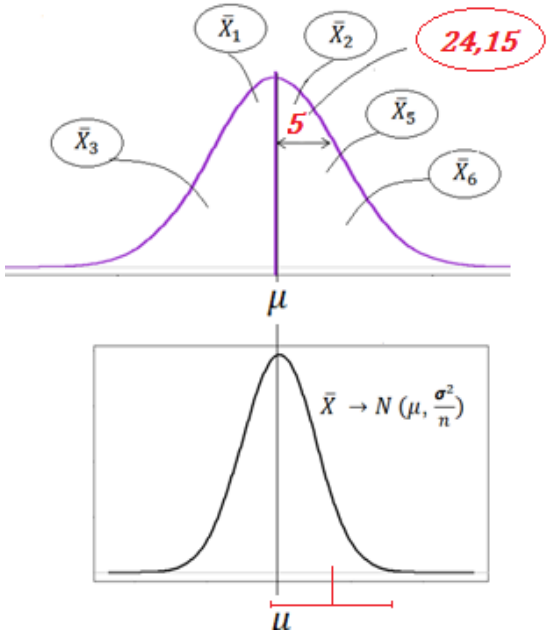
x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
−0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
−0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
−0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
−0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
−0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
−0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
−0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
−0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
−0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
−0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
−1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
−1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
−1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
−1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
−1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
−1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
−1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
−1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
−1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
−1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
−2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
−2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143

Решение

$$\overline{X} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

$$24{,}15 \pm \left(1{,}96 * \frac{5}{\sqrt{100}}\right)$$

[23,17; 25,13]

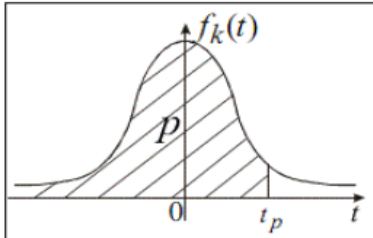


Доверительный интервал. Сигма совокупности неизвестна

Дана выборка, состоящая из роста 10 человек. Оценить средний рост в данной популяции с помощью 95% интервала.

$$\overline{X} \pm t_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

```
import numpy as np
a = np.array([178, 184, 149, 193, 186, 173, 169, 175, 159, 174])

x_1 = np.mean(a) # найдем среднее арифметическое для выборки a
x_1
174.0

D1 = np.var(a, ddof=1)   # несмещенная дисперсия для выборки 1
D1
166.44

# доверительный интервал для среднего
t1 = stats.t.ppf (0.975, 9)
t1
2.2621571627409915

x_1 - t1 * np.sqrt(D/10)
166.05914451355463

x_1 + t1 * np.sqrt (D/10)
181.94085548644537

166.06 ; 181.94
```



Число степеней свободы k	Вероятность, p					
	0,9	**0,95**	**0,975**	**0,99**	**0,995**	**0,9995**
1	2	3	4	5	6	7
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
12	1,356	1,782	2,179	2,681	3,055	4,318
14	1,345	1,761	2,145	2,625	2,977	4,140

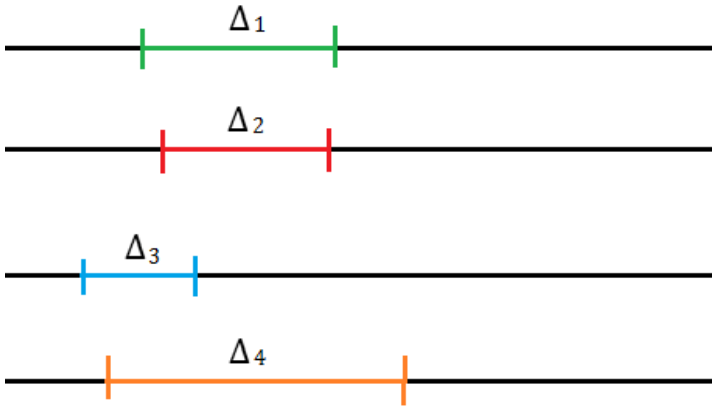
Интервальная оценка для разности средних арифметических

$\Delta \pm t_{\frac{\alpha}{2}} * S_{\Delta}$, где $\Delta = \overline{X_1} - \overline{X_2}$

$D = \frac{1}{2}(D_1 + D_2)$, где D_1 и D_2 – дисперсии обеих групп

$S_{\Delta} = \sqrt{\frac{D}{n_1} + \frac{D}{n_2}}$, где n_1 и n_2 – объемы выборок.

$df = 2 * (n - 1)$, где n- объем выборки



Интервальная оценка для разности средних арифметических

Оценить различие в росте между двумя средними арифметическими популяции a и b с помощью 95% доверительного интервала.

```
import numpy as np
a = np.array([178, 184, 149, 193, 186, 173, 169, 175, 159, 174])

array([178, 184, 149, 193, 186, 173, 169, 175, 159, 174])

b= np.array ([ 150, 154, 167, 165, 171, 150, 158, 170, 175, 160])

len(a)
10

len(b)
10

x_1 = np.mean(a) # найдем среднее арифметическое для выборки a
x_1
174.0

x_2 = np.mean(b) #  среднее для выборки b
x_2
162.0

delta = x_1 -x_2 # разность средних
delta
12.0
```

$$\Delta \pm t_{\frac{\alpha}{2}} * S_{\Delta}$$

```
D1 = np.var(a, ddof=1)   # несмщенная дисперсия для выборки 1
D1
166.44

D2 = np.var(b, ddof=1)  # несмещенная дисперсия для выборки 2
D2
80.0
D = (D1 + D2)/2 #  объединенная оценка дисперсии
D
123.22

SE= np.sqrt (D/10+D/10)  # стандартная ошибка разности средних
SE
4.964317117635058

import scipy.stats as stats
t = stats.t.ppf(0.975, 18) # степени свободы 2*(n-1) = 2*(10 -1) = 18
t
2.10092204024096
```

$$\Delta \pm t\frac{\alpha}{2} * S_{\Delta}$$

$$D = \frac{1}{2}(D_1 + D_2),$$

$$S_{\Delta} = \sqrt{\frac{D}{n_1} + \frac{D}{n_2}}$$

```
L = delta - t*SE # нижняя граница интервала
L
0.7699344742241419

U = delta +t* SE # верхняя граница интервала
U
23.230065525775856

0.77 ; 23.23
```

Доверительный интервал для доли. Стандартное отклонение.

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

В сфере образования страны X работают 20% женщин и 80% мужчин. Рассчитать стандартное отклонение для доли p = 0.2

$\sigma = \sqrt{0.2\ (1-0.2)} = 0.4$

А что будет, если p = 0.5?

$\sigma = \sqrt{0.5\ (1-0.5)} = 0.5,$

для p = 0 (в коллективе вообще нет женщин, одни мужчины)

$\sigma = \sqrt{0\ (1-0)} = 0$

для p = 1 (коллективе все женщины и совсем нет мужчин)

$\sigma = \sqrt{1\ (1-1)} = 0$

Стандартная ошибка для доли

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

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

$\frac{1*2+0*8}{10} = 0.2$, т.е. $\mu = p$

Этим приближением мы пользуемся, когда $n * \hat{p}$ и $n * (1-\hat{p})$ больше 5. Это утверждение нарушается при маленьких объемах выборки и p близких к 0 или 1.

Построение доверительного интервала для доли. Задача ГГГГГГГГГГГГГГГ

Построить 95% доверительный интервал для истинного p – доли носителей некоторого гена. Объем выборки 100. Доля носителей в выборке составляет 0.2
100 * 0.2 = 20 и 100 * (1-0.2) = 80 , оба значения > 5

$$p \pm Z\frac{\alpha}{2} * SE$$

0.2 $\pm$1.96* $\sqrt{\frac{0.2(1-0.2)}{100}}$ = 0.2 $\pm$ 0.08

[0.1216; 0.2784]

В процентах:

[12,16% ; 27,84%]

Маленькие объемы выборок

Если не выполняются условия $n * \hat{p} > 5$ и $n * (1 - \hat{p}) > 5$, используют биномиальное распределение.

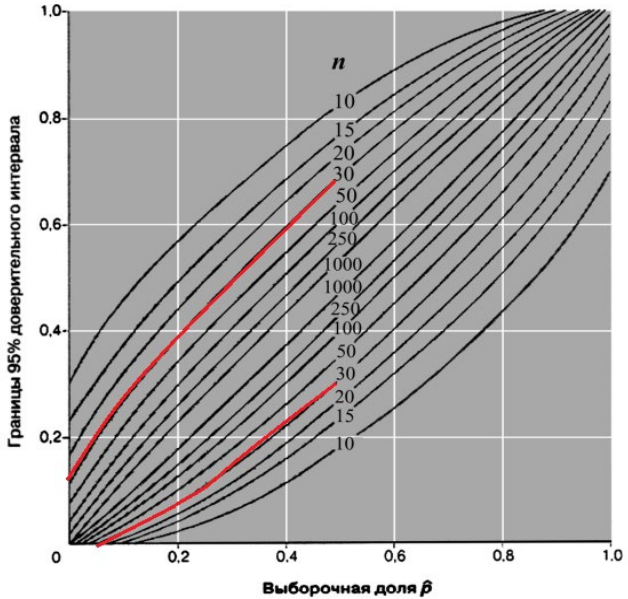
формула Бернулли:

$$C_n^k * p^k * q^{n-k}$$

графический метод:
ГГ
Ни у одного из 30 пациентов препарат не вызвал побочного эффекта. Оценить истинную долю пациентов в популяции, у которых препарат вызовет побочный эффект.

$$\hat{p} = \frac{0}{30} = 0$$

от 0 до 0.13
в процентах от 0% до 13%



Пример из книги «Медико - биологическая статистика» С.Гланц

Задача

У 1 из 5 пациентов, принимавших участие в клинических исследованиях, препарат вызвал побочный эффект. Оценить истинное значение доли пациентов в популяции, у которых препарат вызовет побочный эффект.

$n * p = 5 * 0.2 = 1$

1<5

$C_n^k * p^k * q^{n-k}$

Значение СВ	Доля в выборке	Вероятность	Накопленная вероятность
0	0	0.32768	0.32768
1	0.2	0.4096	0.73728
2	0.4	0.2048	0.94208
3	0.6	0.0512	0.99328
4	0.8	0.0064	0.99968
5	1	0.00032	1

0.32768 +0.4096+ 0.2048 = 0.94208

94% интервал составляет 0 – 0.4

Сравнение долей

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{где } p = \frac{m_1 + m_2}{n_1 + n_2},$$

Критерий Z с поправкой Йейтса на непрерывность:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}(\frac{1}{n_1} + \frac{1}{n_2})}{\sqrt{p(1-p) * (\frac{1}{n_1} + \frac{1}{n_2})}}$$

Задача. Сравнение долей.

Есть две группы студентов объемом 56 и 61, которые сдают международный экзамен на знание иностранного языка. Максимальное число баллов за тест 120. Высокая оценка считается выше 100 баллов. В первой группе высокую оценку получили 7 студентов, а во второй 22. Есть ли статистически значимые различия в долях студентов этих двух групп, сдавших тест на высокий балл.

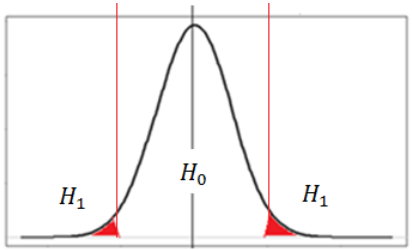
Решение:

$n_1 = 56 \quad m_1 = 7 \quad \hat{p}_1 = \frac{7}{56}$

$n_2 = 61 \quad m_2 = 22 \quad \hat{p}_2 = \frac{22}{61}$

$\hat{p} = \frac{7+22}{56+61} = \frac{29}{117}$

$$z = \frac{\left|\frac{7}{56} - \frac{22}{61}\right| - \frac{1}{2}\left(\frac{1}{56} + \frac{1}{61}\right)}{\sqrt{\frac{29}{117}\left(1 - \frac{29}{117}\right) * \left(\frac{1}{56} + \frac{1}{61}\right)}} \approx 2.725$$



$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p = \frac{m_1 + m_2}{n_1 + n_2}$$

$$z = \frac{\left|\hat{p}_1 - \hat{p}_2\right| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{p(1-p) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Интервал для разности долей

На одном сайте из 153 посетителей 75 оформили заказ, а на другом сайте из 120 посетителей заказ оформили 50 человек. Оценить с помощью доверительного интервала разность долей покупателей, совершивших покупку.

$\widehat{p_1} = \frac{75}{153} = 0.490,$

$\widehat{p_2} = \frac{50}{120} = 0.417$

$\Delta = 0.490 - 0.417 = 0.073$

$S_\Delta = \sqrt{p_{\text{общ}}(1 - p_{\text{общ}})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} =$

$= \sqrt{0.458 * (1 - 0.458)\left(\frac{1}{153} + \frac{1}{120}\right)} = 0.06$

$0.073 \pm 1.96 * 0.06 \Rightarrow [-0.045;\ 0.009]$

$\Delta \pm z_{a/2} * S_\Delta$, где Δ – разность долей

S_Δ – стандартная ошибка разности долей

$$S_\Delta = \sqrt{p_{\text{общ}}\left(1 - p_{\text{общ}}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

где $p_{\text{общ}} = \frac{m_1 + m_2}{n_1 + n_2}$

Конец