Подготовка к контрольной работе

Теория информации

$$H(X) = -\sum_i p_i \log_2(p_i)$$

$$H(X) = -\int_a^b f(x)\log_2(f(x))dx$$

$$H(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$CE(X||Y) = -\sum p(x)\log_2 q(x)$$

$$H(X|Y) = H(X,Y) - H(Y)$$

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$D_{KL}(X\|Y) = CE(X\|Y) - H(X) = \sum_{x\in X} p(x)\log_2\frac{p(x)}{p(y)}$$

Множественная проверка гипотез

	H_0 не отвергается	H_0 отвергается	Итого
H_0 верна	V	U	m_0
H_0 неверна	S	T	$m-m_0$
Итого	R	$m-R$	m

- FWER $= P\{U>0\} = E[I\{U>0\}] = \frac{m_0\cdot\alpha}{m}$
- FDR $= E[\frac{U}{max\{U+T,1\}}] =$
- поправка Бонферрони $\alpha_{ind} = \frac{\alpha}{m}$
- поправка Холма-Бонферрони $\alpha_k = \frac{\alpha}{m+1-k}$
- процедура Бенджамини-Хокберга $\alpha_k = \frac{\alpha\cdot k}{m}$

1.1 EM-алгоритм (ДЗ1 2022, №5)

Для $i \in \{1, \dots, 7\}$ пусть Y_i – случайная величина, обозначающая логарифм количества мёда в i-м дереве; Z_i – случайная величина, равная 0, если i-е дерево хорошее, и 1, если плохое; y_i – реальное наблюдение логарифма количества мёда в i-м дереве.

В этой задаче вектор параметров $\theta = \begin{bmatrix} \mu_g & \mu_b \end{bmatrix}^\top$.

E-шаг

$$Q(\theta \mid \theta_{\text{old}}) = \mathbb{E}_{\theta_{\text{old}}}[\ell_\theta(Y, Z) \mid Y = y] = \sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \ell_\theta(y, z) \cdot p_{\theta_{\text{old}}}(Z = z \mid Y = y)$$

(Сверху сумма по всем векторам z длины 7 из нулей и единиц, в которых ровно 2 единицы.)

Посчитаем первый множитель (зависящий от θ):

$$\ell_\theta(y, z) = \ln p_\theta(y \mid Z = z) + \ln p_\theta(Z = z)$$

Вероятность $p_\theta(Z = z)$ для каждого z одинакова и равна $1/C_7^2$. Распишем первое слагаемое:

$$\ln p_\theta(y \mid Z = z) = \ln \prod_{i=1}^{7} p_\theta(y_i \mid Z = z) = \ln \prod_{i=1}^{7} p_\theta(y_i \mid Z_i = z_i)$$

$$= \sum_{i=1}^{7} \ln p_\theta(y_i \mid Z_i = z_i) = \sum_{i=1}^{7} \ln \left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mu_{z_i})^2}{2} \right) \right] = -\frac{7}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{7} (y_i - \mu_{z_i})^2$$

Здесь $\mu_{z_i} = \mu_g \cdot [z_i = 0] + \mu_b \cdot [z_i = 1]$, где μ_g и μ_b оба взяты из θ.

Посчитаем второй множитель (зависящий от θ_{old}):

$$p_{\theta_{\text{old}}}(Z = z \mid Y = y) = \{\text{Байес}\} = \frac{p_{\theta_{\text{old}}}(y \mid Z = z) \cdot p_{\theta_{\text{old}}}(Z = z)}{p_{\theta_{\text{old}}}(y)}$$

$$= \{\text{формула полной вероятности}\} = \frac{p_{\theta_{\text{old}}}(y \mid Z = z) \cdot p_{\theta_{\text{old}}}(Z = z)}{\sum\limits_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} p_{\theta_{\text{old}}}(y \mid Z = z) \cdot p_{\theta_{\text{old}}}(Z = z)}$$

Вероятность $p_{\theta_{\text{old}}}(Z = z)$ для каждого z одинакова (и равна $1/C_7^2$), поэтому она сократится в числителе и знаменателе. А вероятность $p_{\theta_{\text{old}}}(y \mid Z = z)$ считается аналогично тому, как мы считали $\ln p_\theta(y \mid Z = z)$ выше, только без логарифма и уже используя не θ, а θ_{old}:

$$p_{\theta_{\text{old}}}(y \mid Z = z) = \prod_{i=1}^{7} \left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mu_{z_i}^{\text{old}})^2}{2} \right) \right]$$

Здесь $\mu_{z_i}^{\text{old}} = \mu_g \cdot [z_i = 0] + \mu_b \cdot [z_i = 1]$, где μ_g и μ_b оба взяты из θ_{old}.

Осталось подставить всё это выше, и мы получим функцию $Q(\theta \mid \theta_{\text{old}})$.

M-шаг

Хотим найти новую оценку максимального правдоподобия для μ_g и μ_b. Я буду искать новую оценку для $\mu_k = \mu_g \cdot [k=0] + \mu_b \cdot [k=1]$ для произвольного $k \in \{0,1\}$, чтобы убить двух ежей сразу.

Продифференцируем $Q(\theta \mid \theta_{\text{old}})$ по μ_k:

$$\frac{d}{d\mu_k} Q(\theta \mid \theta_{\text{old}}) = -\frac{1}{2}\frac{d}{d\mu_k} \sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k](y_i - \mu_k)^2 \right)$$

$$= \sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k](y_i - \mu_k) \right)$$

Приравняем производную к 0:

$$\sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k](y_i - \mu_k) \right) = 0$$

$$\sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k] y_i \right) = \mu_k \cdot \sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k] \right)$$

Последняя сумма в правой части всегда равна 5, если $k = 0$, и 2, если $k = 1$. Обозначим $c_k = 5[k=0] + 2[k=1]$ и вынесем её влево:

$$\sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k] y_i \right) = \mu_k \, c_k \cdot \sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} p_{\theta_{\text{old}}}(Z = z \mid Y = y)$$

Теперь очевидно, что сумма в правой части просто равна 1, так как это сумма вероятностей всех возможных значений вектора Z.

$$\sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k] y_i \right) = \mu_k \, c_k$$

$$\mu_k = \frac{1}{c_k} \sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = k] y_i \right)$$

Таким образом, новый вектор θ будет равен

$$\theta = \begin{bmatrix} \mu_g \\ \mu_b \end{bmatrix} = \begin{bmatrix} \dfrac{1}{5} \displaystyle\sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = 0] y_i \right) \\ \dfrac{1}{2} \displaystyle\sum_{\substack{z \in \{0,1\}^7 \\ \sum z_i = 2}} \left(p_{\theta_{\text{old}}}(Z = z \mid Y = y) \sum_{i=1}^{7} [z_i = 1] y_i \right) \end{bmatrix},$$

где $p_{\theta_{\text{old}}}(Z = z \mid Y = y)$ мы уже посчитали выше.

1.2 LR,LM,W и точечный ММП (ДЗ1 2021, №1)

Компания «Напиши-ка» производит три вида ручек: синие, красные и зелёные. Глава аналитического отдела компании Данил хочет понять, какая из ручек скорее всего «выстрелит», а какая не будет пользоваться успехом у покупателей. Для этого он анализирует выборку в 300 проданных ручек. Оказалось, что из них 150 синих, 100 красных и 50 зелёных ручек. Данил уверен, что ручки продаются независимо друг от друга, и вероятность того, что будет продана синяя ручка, равна p_1, а что красная p_2.

Ручка	С	К	З
N	150	100	50
P	p_1	p_2	$1 - p_1 - p_2$

1. Обозначим $p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$. Найдите $\hat{p}_{ML}$ интуитивно, не выписывая правдоподобие, и поясните, как вы это сделали.

   В силу независимости продажи ручек кажется очевидным предположить, что вероятности купить ручку сообразны долям уже проданных ручек исходя из имеющейся выборки, то есть $\hat{p_1} = 0.5, \hat{p_2} = \frac{1}{3}$

2. Выпишите функцию правдоподобия и найдите $\hat{p}_{ML}$ как точку её глобального максимума.

   $L(X|p) = p_1^{150} \cdot p_2^{100} \cdot (1 - p_1 - p_2)^{50}$

   $l = \ln L = 150 \ln p_1 + 100 \ln p_2 + 50 \ln(1 - p_1 - p_2) \to max_{p1,p2}$

   $l'_{p1} = \frac{150}{p1} - \frac{50}{1-p_1-p_2}$

   $\hat{p}_2 = \frac{2-2p_1}{3}$

   $l'_{p2} = \frac{100}{p2} - \frac{50}{1-p_1-p_2}$

   $4 \cdot \hat{p}_1 = 3 - 3\frac{2-2p_1}{3}$

   $\hat{p}_1 = \frac{1}{2}, \hat{p}_2 = \frac{2-1}{3} = \frac{1}{3}$

   Проверим, что найденные оценки действительно максимизируют функцию правдоподобия:

   $l''_{p_1,p_1} = \frac{-150}{p_1^2} - \frac{50}{(1-p_1-p_2)^2}$

   $l''_{p_2,p_2} = \frac{-100}{p_2^2} - \frac{50}{(1-p_1-p_2)^2}$

   $l''_{p_1,p_2} = -\frac{50}{(1-p_1-p_2)^2}$

   Получаем такой гессиан:

   $H(\hat{p}_{ML}) = \begin{pmatrix} -2400 & -1800 \\ -1800 & -2700 \end{pmatrix}$

   $\Delta_1 = -2400 \ \Delta_2 = 2400 \cdot 2700 - 1800^2 > 0$

   В соответствии с критерием Сильвестра имеет отрицательно определённую матрицу, что свидетельствует об успешном нахождении максимума функции правдоподобия, который полностью совпадает с интуитивной оценкой, данной в предыдущем пункте

   Заодно найдём оценку дисперсии оценок параметров:

   $$Var(\hat{p}_{ML}) = \hat{I}(\hat{p}_{ML})^{-1} = [E(-H)]^{-1} = \begin{pmatrix} 2400 & 1800 \\ 1800 & 2700 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{1200} & -\frac{1}{1800} \\ -\frac{1}{1800} & \frac{1}{1350} \end{pmatrix}$$

3. Проверьте гипотезу

   $$\begin{cases} H_0 : p_1 = 0.2, \\ H_A : p_1 \neq 0.2 \end{cases}$$

на уровне значимости 5% при помощи тестов LR и LM.

$LR = 2(max_{p_1,p_2} l(p_1,p_2) - max_{p_2} l(p_1,p_2))$

$max_{p_1,p_2} l(p_1,p_2) = l(\hat{p}_1,\hat{p}_2) = 150\ln 0.5 + 100\ln\frac{1}{3} + 50\ln\frac{1}{6} = -303.42$

$max_{p_2} l(p_1 = 0.2, p_2) = l(p_1 = 0.2, \hat{p}_2) = 150\ln 0.2 + 100\ln\frac{1}{3} + 50\ln(0.8 - \frac{1}{3}) = -389.384$

Note:p_2 надо максимизировать заново!

$LR = 171.928$

Даже без подбора критического значения для LR-теста очевидно, что p-value нулевой гипотезы примерно равно 0, следовательно она отвергается на любом разумном уровне значимости

$LM = \hat{s}(p_1 = 0.2)^T \hat{I}_F(p_1 = 0.2)^{-1} \cdot \hat{s}(p_1 = 0.2)$

$\hat{s}(p_1 = 0.2) = \begin{pmatrix} \frac{150}{p1} - \frac{50}{1-p_1-p_2} \\ \frac{100}{p2} - \frac{50}{1-p_1-p_2} \end{pmatrix} = \begin{pmatrix} 642.857 \\ 192.857 \end{pmatrix}$

Находим информацию Фишера при условии новых ограничений!

$\hat{I}_F(p_1 = 0.2) = E(-H) = \begin{pmatrix} 3979.36 & 229.36 \\ 229.358 & 1129.36 \end{pmatrix}$

$\hat{I}_F(p_1 = 0.2)^{-1} = \frac{1}{3979.36 \cdot 1129.36 - 229.36^2} \begin{pmatrix} 1129.36 & -229.36 \\ -229.358 & 3979.36 \end{pmatrix}$

Агрегируя получившиеся значение, можем получить наблюдаемое значение LM-статистики - $LM = 125.601$.

В случае LM, так же, как и в случае с LR-теста, нулевая гипотеза $p_1 = 0.2$ отвергается на любом разумном уровне значимости.

4. Проверьте гипотезу

$$\begin{cases} H_0 : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix}, \\ H_A : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \neq \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix} \end{cases}$$

на уровне значимости 5% при помощи тестов LR и W.

$LR = 2(max_{p_1,p_2} l(p_1,p_2) - l(p_1 = 0.3, p_2 = 0.2))$

$max_{p_1,p_2} l(p_1,p_2) = *\textbf{узнали ранее}* = -303.42$

$l(p_1 = 0.3, p_2 = 0.2) = l(p_1 = 0.2, \hat{p}_2) = 150\ln 0.3 + 100\ln 0.2 + 50\ln(0.5) = -376.197$

$LR = 2(-303.42 + 376.2) = 145.56$ Формально проверим нашу гипотезу:

- p-value(LR)= $2 \cdot min\{P\{z \leq 145.56|H_0\}, P\{z \geq 145.56|H_0\}\} < 0.00001$
- **p-value** $< \alpha = 0.05 \to H_0$ отвергается
- $\chi^2_{l_c r} = \chi^2_{\alpha=0.025,d=2} = 0.0506, \chi^2_{l_c r} = \chi^2_{\alpha=0.975,d=2} = 7.3778 \to \chi^2_{obs}$ не входит в доверительный интервал, позволяющий не отвергнуть гипотезу

$W = (\hat{\gamma}_{ML} - \gamma_0)^T Var(\hat{\gamma}_{ML})^{-1}(\hat{\gamma}_{ML} - \gamma_0) = \begin{pmatrix} (0.5 - 0.3) & (\frac{1}{3} - 0.2) \end{pmatrix} \begin{pmatrix} \frac{1}{1200} & -\frac{1}{1800} \\ -\frac{1}{1800} & \frac{1}{1350} \end{pmatrix} \begin{pmatrix} (0.5 - 0.3) \\ (\frac{1}{3} - 0.2) \end{pmatrix} =$

В Вальде ничего не надо максимизировать заново!

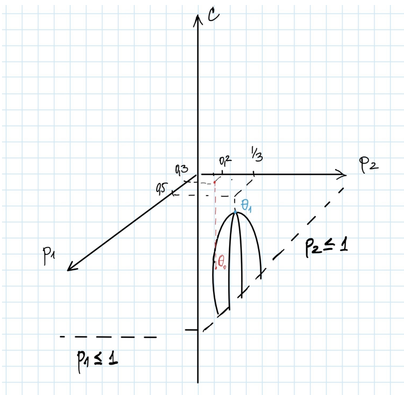


Рис. 1.1: — График логарифма правдоподобия в трёхмерной плоскости

5. Постройте график логарифма правдоподобия в трёхмерной плоскости. Покажите на графике $\hat{p}_{ML}$ визуальную интерпретацию тестов LR и W для гипотезы из предыдущего пункта.

6. Постройте 95%-ый доверительный интервал для p_3.

7. Постройте 99%-ый доверительный интервал для $p_1 + p_2$.

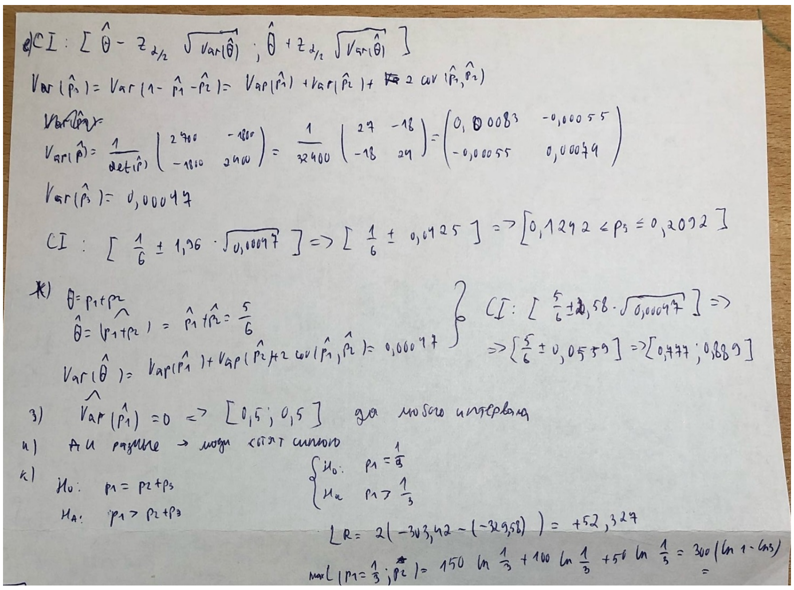
8. Постройте 90%-ый доверительный интервал для $\hat{p}_1$.

   Подсказка: помните, что мы работаем в рамках частотного подхода.

9. Приведите разумное интерпретируемое определение того, что ручка «выстрелила».

10. Пользуясь определением из предыдущего пункта, сформулируйте гипотезу о том, что «выстрелит» ручка синего цвета и проверьте её при помощи любого из тестов LR, LM или W на уровне значимости 5%.

    Решение оставшихся пунктов приведено на картинке ниже:



е) $CI: [\hat\theta - z_{\alpha/2}\sqrt{Var(\hat\theta)}; \hat\theta + z_{\alpha/2}\sqrt{Var(\hat\theta)}]$

$Var(\hat p_3) = Var(1 - \hat p_1 - \hat p_2) = Var(\hat p_1) + Var(\hat p_2) + 2\,cov(\hat p_1, \hat p_2)$

$Var(\hat p) = \frac{1}{\det(I)}\begin{pmatrix} 2700 & -1800 \\ -1800 & 2400 \end{pmatrix} = \frac{1}{32400}\begin{pmatrix} 27 & -18 \\ -18 & 24 \end{pmatrix} = \begin{pmatrix} 0,00083 & -0,00055 \\ -0,00055 & 0,00074 \end{pmatrix}$

$Var(\hat p_3) = 0,00047$

$CI: [\frac{1}{6} \pm 1,96 \cdot \sqrt{0,00047}] \Rightarrow [\frac{1}{6} \pm 0,0425] \Rightarrow [0,1242 \leq p_3 \leq 0,2092]$

ж) $\theta = p_1 + p_2$

$\hat\theta = \widehat{(p_1+p_2)} = \hat p_1 + \hat p_2 = \frac{5}{6}$

$Var(\hat\theta) = Var(\hat p_1) + Var(\hat p_2) + 2\,cov(\hat p_1, \hat p_2) = 0,00047$

$CI: [\frac{5}{6} \pm 2,58 \cdot \sqrt{0,00047}] \Rightarrow [\frac{5}{6} \pm 0,0559] \Rightarrow [0,777; 0,889]$

з) $\widehat{Var}(\hat p_1) = 0 \Rightarrow [0,5; 0,5]$ для любого интервала

и) А и разные → люди хотят синюю

к) $H_0: p_1 = p_2 + p_3$
$H_A: p_1 > p_2 + p_3$

$\begin{cases} H_0: & p_1 = \frac{1}{3} \\ H_a: & p_1 > \frac{1}{3} \end{cases}$

$LR = 2(-303,42 - (-329,58)) = +52,327$

$\max L(p_1 = \frac{1}{3}; \hat p_2) = 150 \ln\frac{1}{3} + 100 \ln\frac{1}{3} + 50 \ln\frac{1}{3} = 300(\ln 1 - \ln 3) =$

Рис. 1.2: — Решение второй части задачи 1

1.3 Множественная проверка гипотез (ДЗ1 2022, №6)

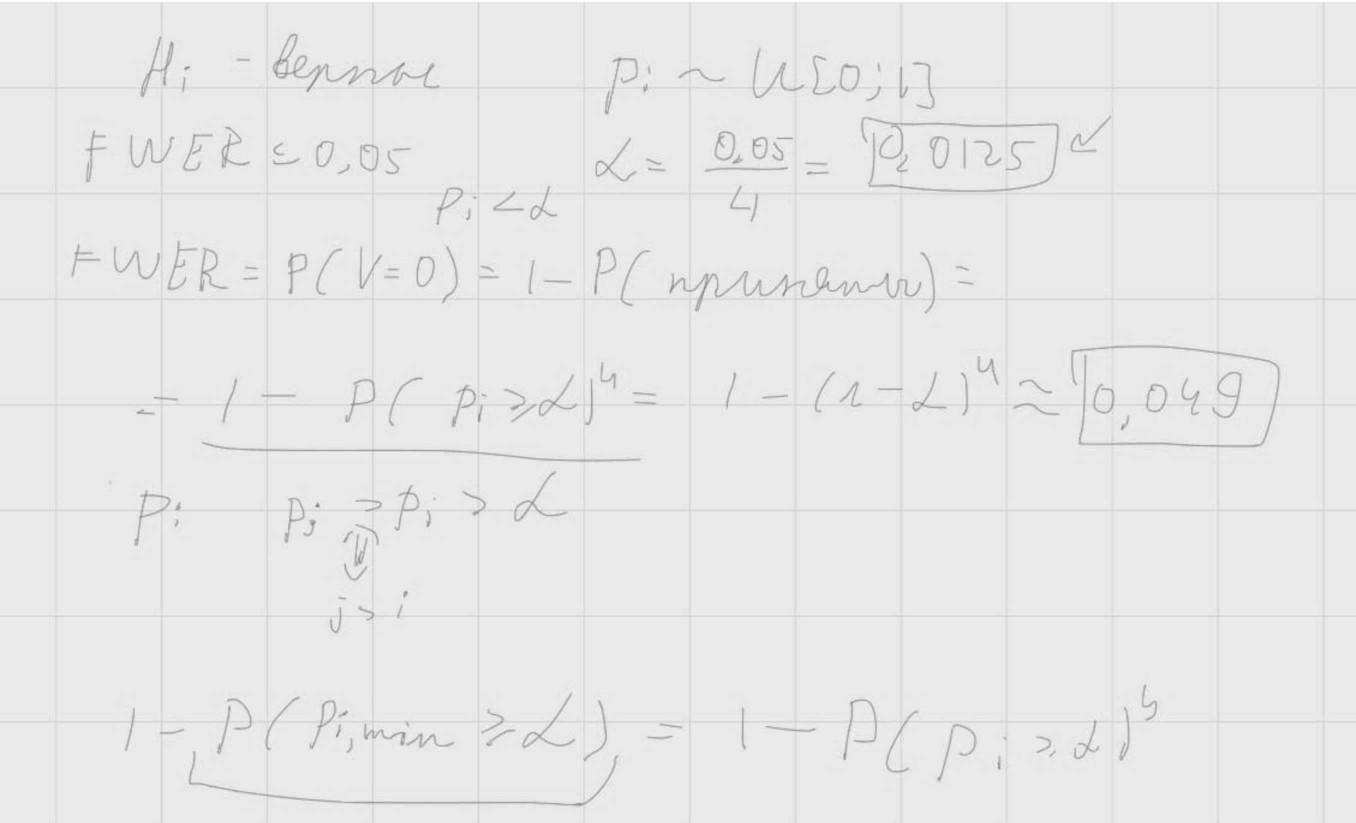


Рис. 1.3: Решение от Артёма Беляева