Задача 5

а) Пусть $odds_+[i]$ – шанс написать экзамен на хорошую оценку в i-й группе, $i \in \{C, V\}$, где C – согласные, V – гласные.

Тогда $\widehat{odds_+[i]} = \frac{n_i^+}{n_i^-} = \frac{n_i^+}{n_i} : \frac{n_i^-}{n_i} = \frac{\hat{p}_i}{1-\hat{p}_i} \Rightarrow \widehat{OR} = \frac{\widehat{odds_+[C]}}{\widehat{odds_+[V]}} = \frac{\hat{p}_V}{1-\hat{p}_V} : \frac{\hat{p}_c}{1-\hat{p}_c} \Rightarrow$

$\Rightarrow \left\| \text{т.к. } \hat{p}_c = \bar{C},\ \hat{p}_V = \bar{V}, \text{ то } \widehat{odds_+[C]} = f(\bar{C}),\ \widehat{odds_+[V]} = f(\bar{V}) \right\| \Rightarrow$

$\Rightarrow \ln \widehat{OR} = \ln f(\bar{V}) - \ln f(\bar{C}) = \ln \frac{\hat{p}_V}{1-\hat{p}_V} - \ln \frac{\hat{p}_c}{1-\hat{p}_c}.$

Рассмотрим $\ln \frac{\hat{p}}{1-\hat{p}} \overset{\text{Д.М.}}{\approx} \ln \frac{p}{1-p} + \frac{1-p-p\cdot(-1)}{(1-p)^2} \cdot \frac{1-p}{p}(\hat{p}-p) \Leftrightarrow$

$\Leftrightarrow \ln \frac{\hat{p}}{1-\hat{p}} \approx \ln \frac{p}{1-p} + \frac{1}{(1-p)p}(\hat{p}-p) \Rightarrow E\left(\ln \frac{\hat{p}}{1-\hat{p}}\right) = \ln \frac{p}{1-p},\ Var\left(\ln \frac{\hat{p}}{1-\hat{p}}\right) =$

$= \frac{1}{(1-p)^2 \cdot p^2} \cdot Var(\hat{p}) = \frac{1}{(1-p)^2 p^2} \cdot \frac{Var(\overset{\sim Bi(1,p)}{X_i})}{n} = \frac{1}{(1-p)pn}.$

Тогда $\ln \widehat{OR} \overset{as.}{\sim} \mathcal{N}\left(\ln \frac{p_V}{1-p_V} - \ln \frac{p_c}{1-p_c}, \frac{1}{(1-p_V)p_V \cdot n_V} + \frac{1}{(1-p_c)p_V \cdot n_c}\right) \Leftrightarrow$

$\Leftrightarrow \ln \widehat{OR} \overset{as.}{\sim} \mathcal{N}\left(\ln OR, \frac{1}{(1-p_V)p_V \cdot n_V} + \frac{1}{(1-p_c)\cdot p_c \cdot n_c}\right) \Rightarrow$

$\Rightarrow \frac{\ln \widehat{OR} - \ln OR}{se(\ln \widehat{OR})} \overset{as.}{\sim} \mathcal{N}(0,1) \Rightarrow P\left(-z_{1-\frac{\alpha}{2}} \le \frac{\ln \widehat{OR} - \ln OR}{se(\ln \widehat{OR})} \le z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$

$\Leftrightarrow P\left(\ln \widehat{OR} - z_{1-\frac{\alpha}{2}} \cdot se(\ln \widehat{OR}) \le \ln OR \le \ln \widehat{OR} + z_{1-\frac{\alpha}{2}} \cdot se(\ln \widehat{OR})\right) = 1-\alpha \Leftrightarrow$

$\Leftrightarrow P\left(\exp\left(\ln \widehat{OR} - z_{1-\frac{\alpha}{2}} \cdot se(\ln \widehat{OR})\right) \le OR \le \exp\left(\ln \widehat{OR} + z_{1-\frac{\alpha}{2}} \cdot se(\ln \widehat{OR})\right)\right) = 1-\alpha.$

$\Downarrow$ код

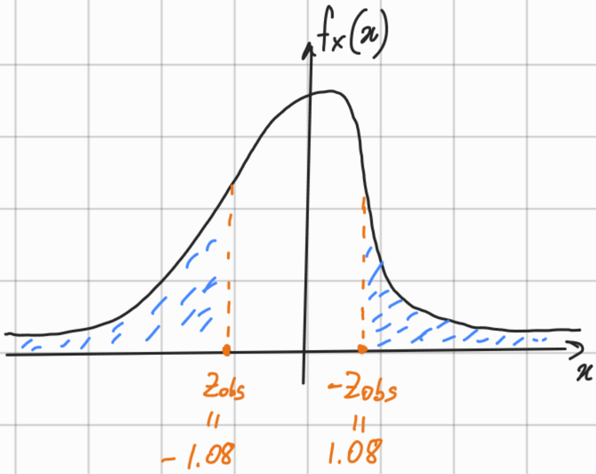
$OR \in [0.39; 1.32]$

Теперь перейдем к проверке гипотезы.

$$H_0: OR = 1,\quad H_1: OR \neq 1 \quad \sim \quad H_0: \ln OR = 0,\quad H_1: \ln OR \neq 0 \ \Big| \ \overset{H_0}{\Rightarrow} z_{obs} = \frac{\ln \hat{OR} - 0}{se(\ln \hat{OR})} \overset{ac.}{\sim} \mathcal{N}(0,1) \overset{код}{\Rightarrow}$$

$$\overset{код}{\Rightarrow} z_{obs} = -1.08 \Rightarrow |z_{obs}| < z_{crit} \Rightarrow H_0 \text{ не отвергается.}$$

Найдем $p\text{-value} = 2\cdot\Phi(z_{obs}) = 2\cdot\Phi(-1.08) = 2\cdot 0.14 = 0.28$



б) Рассмотрим отношение вероятностей $\frac{p_v}{p_c}$. Перейдем к $\ln\frac{p_v}{p_c} = \ln p_v - \ln p_c$

Заметим, что $\ln\hat{p} \overset{Д.М.}{\approx} \ln p + \frac{1}{p}(\hat{p} - p) \Rightarrow E(\ln\hat{p}) = \ln p,\ Var(\ln\hat{p}) = \frac{1}{p^2} Var(\hat{p}) =$

$$= \frac{1}{p^2}\cdot\frac{Var(\overset{\sim Bi(1,p)}{X_i})}{n} = \frac{p(1-p)}{p^2 n} = \frac{1-p}{pn},\ \text{и}\ \ln\hat{p} \overset{ac.}{\sim} \mathcal{N}\left(\ln p, \frac{1-p}{pn}\right).$$

Тогда $\ln\hat{p}_v - \ln\hat{p}_c \overset{ac.}{\sim} \mathcal{N}\left(\ln p_v - \ln p_c, \frac{1-p_c}{p_c n_c} + \frac{1-p_v}{p_v n_v}\right) \Rightarrow$

$$\Rightarrow P\left(-z_{crit} \leq \frac{\ln\hat{p}_v - \ln\hat{p}_c - (\ln p_v - \ln p_c)}{se(\ln\hat{p}_v - \ln\hat{p}_c)} \leq z_{crit}\right) = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(\ln\hat{p}_v - \ln\hat{p}_c - z_{crit}\cdot se(\ln\hat{p}_v - \ln\hat{p}_c) \leq \ln\frac{p_v}{p_c} \leq \ln\hat{p}_v - \ln\hat{p}_c + z_{crit}\cdot se(\ln\hat{p}_v - \ln\hat{p}_c)\right) = 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(\exp(\ln\hat{p}_v - \ln\hat{p}_c - z_{crit}\cdot se(\ln\hat{p}_v - \ln\hat{p}_c)) \leq \frac{p_v}{p_c} \leq \exp(\ln\hat{p}_v - \ln\hat{p}_c + z_{crit}\cdot se(\ln\hat{p}_v - \ln\hat{p}_c)\right) = 1-\alpha \overset{код}{\Rightarrow}$$

$$\overset{код}{\Rightarrow} \frac{p_v}{p_c} \in [0.6; 1.18]$$

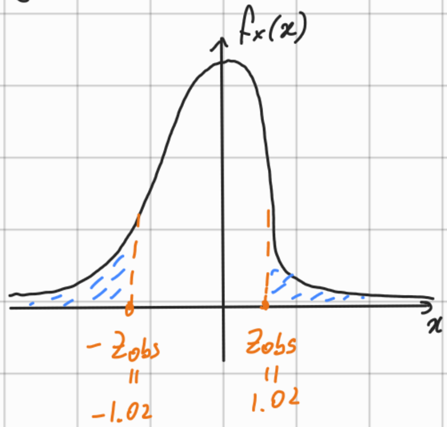
Теперь проверим гипотезу.

$$H_0: \frac{p_v}{p_c} = 1 \sim H_0: \ln\frac{p_v}{p_c} = 0 \qquad \overset{H_0}{\Rightarrow} z_{obs} = \frac{\ln\hat{p}_v - \ln\hat{p}_c - 0}{se(\ln\hat{p}_v - \ln\hat{p}_c)} \overset{as.}{\sim} N(0,1) \overset{код}{\Rightarrow}$$

$$H_1: \frac{p_v}{p_c} \neq 1 \qquad H_1: \ln\frac{p_v}{p_c} \neq 0$$

$\Rightarrow z_{obs} = -1.02 \Rightarrow |z_{obs}| < z_{crit} \Rightarrow H_0$ не отвергается.

Найдем p-value $= 2\cdot\Phi(z_{obs}) = 2\cdot\Phi(-1.02) = 0.31$



в) Сгенерируем по n_boot бутстрапированных выборок для C и V, где C – базовая выборка баллов студентов, фамилии которых начинаются с согласных, V – с гласных. Тогда для каждой пары C_i^*, V_i^* можем посчитать $\widehat{OR}^*$. Затем отбросим по 2,5% с каждой стороны гистограммы получившегося списка $\{\widehat{OR}_1^*, \dots, \widehat{OR}_{n\_boot}^*\}$, откуда выделим границы искомого Д.И.: $q_l \leq OR \leq q_r$.

Таким образом, $\widehat{OR}^* = \frac{\hat{p}_v^*}{1-\hat{p}_v^*} : \frac{\hat{p}_c^*}{1-\hat{p}_c^*}$

Из кода получим, что $q_l = 0.37$, $q_r = 1.34 \Rightarrow OR \in [0.37, 1.34]$

Теперь рассмотрим $H_0: OR = 1$, $H_1: OR \neq 1 \Rightarrow$ т.к. $1 \in [0.37; 1.34]$, то H_0 не отвергается.