Домашнее задание 2

Задание 1

Пусть $X_1, \ldots, X_n \sim$ Gamma(α, β). Тогда

$\mathrm{E}(X_1) = \alpha\beta$

$\mathrm{E}(X_1^2) = \mathrm{V}X_1 + (\mathrm{E}X_1)^2 = \alpha\beta^2 + \alpha^2\beta^2 = \alpha(\alpha+1)\beta^2$

Согласно методу моментов получаем систему для нахождения нужных оценок:

$$\begin{cases} \overline{X}_n = \hat{\alpha}\hat{\beta} \\ \overline{X^2}_n = \hat{\alpha}(\hat{\alpha}+1)\hat{\beta}^2 \end{cases} \Rightarrow \begin{cases} \dfrac{\overline{X}_n}{\hat{\alpha}} = \hat{\beta} \\ \overline{X^2}_n = \hat{\alpha}(\hat{\alpha}+1)\dfrac{(\overline{X}_n)^2}{\hat{\alpha}^2} \end{cases} \Rightarrow \begin{cases} \dfrac{\overline{X}_n}{\hat{\alpha}} = \hat{\beta} \\ \overline{X^2}_n = (1 + \dfrac{1}{\hat{\alpha}})(\overline{X}_n)^2 \end{cases} \Rightarrow \begin{cases} \dfrac{\overline{X}_n}{\hat{\alpha}} = \hat{\beta} \\ \dfrac{\overline{X^2}_n}{\overline{X}_n^{2}} - 1 = \dfrac{1}{\hat{\alpha}} \end{cases} \Rightarrow$$

$$\begin{cases} \hat{\alpha} = \dfrac{(\overline{X}_n)^2}{\overline{X^2}_n - (\overline{X}_n)^2} \\ \hat{\beta} = \dfrac{\overline{X^2}_n - (\overline{X}_n)^2}{\overline{X}_n} \end{cases}$$

Задание 2

а) Поскольку ОМП не зависит от параметризации, то ОМП для $\psi = p_1 - p_2$ будет равна $\hat{\psi} = \hat{p}_1 - \hat{p}_2$, где $\hat{p}_1$ и $\hat{p}_2$ − оценки максимального правдоподобия для p_1 и p_2 соответственно.

Имеем случайный вектор (X_1, X_2), где X_1 и X_2 независимые с.в. и $X_1 \sim Bin(n_1, p_1)$, $X_2 \sim Bin(n_2, p_2)$. Найдем ОМП p_1 и p_2.

Многомерное распределение вероятности имеет вид $f((x, y), (p_1, p_2)) = C_{n_1}^{x} p_1^{x} (1-p_1)^{n_1 - x} C_{n_2}^{y} p_2^{y} (1-p_2)^{n_2 - y}$, поскольку компоненты вектора независимы.

Функция правдоподобия: $\mathcal{L}_n(p_1, p_2) = f((X_1, X_2), (p_1, p_2)) = C_{n_1}^{X_1} p_1^{X_1} (1-p_1)^{n_1 - X_1} C_{n_2}^{X_2} p_2^{X_2} (1-p_2)^{n_2 - X_2}$

Логарифм функции правдоподобия:

$l_n(p_1, p_2) = \left[\ln(C_{n_1}^{X_1}) + X_1\ln(p_1) + (n_1 - X_1)\ln(1 - p_1)\right] + \left[\ln(C_{n_2}^{X_2}) + X_2\ln(p_2) + (n_2 - X_2)\ln(1 - p_2)\right]$

Заметим, что логарифм функции правдоподобия разбивается на сумму двух функций, каждая из которых зависит лишь от одного p_i. Поэтому можно минимизировать каждую из функций в отдельности. Также учтем, что эти функции симметричны с точностью до замены индексов).

Найдем производные (в формулах ниже $i = 1, 2$):

$$\frac{\partial l_n(p_1, p_2)}{\partial p_i} = \frac{X_i}{p_i} - \frac{n_i - X_i}{1 - p_i} = \frac{X_i - X_i p_i - n_i p_i + X_i p_i}{p_i(1 - p_i)} = \frac{X_i - n_i p_i}{p_i(1 - p_i)}$$

$$\frac{\partial^2 l_n(p_1, p_2)}{\partial p_i^2} = -\frac{X_i}{p_i^2} - \frac{n_i - X_i}{(1 - p_i)^2} < 0$$

Последнее неравенство следует из того, что $0 \le X_i \le n_i$.

Поэтому искомые ОМП $\hat{p}_1 = \frac{X_1}{n_1}$ и $\hat{p}_2 = \frac{X_2}{n_2}$.

Значит, $\hat{\psi} = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$

b)

$$\mathrm{E}\left[\frac{\partial^2 l_n(p_1, p_2)}{\partial p_i^2}\right] = -\mathrm{E}\left[\frac{X_i}{p_i^2} - \frac{n_i - X_i}{(1 - p_i)^2}\right] = -\frac{\mathrm{E}X_i}{p_i^2} - \frac{n_i - \mathrm{E}X_i}{(1 - p_i)^2} = -\frac{n_i p_i}{p_i^2} - \frac{n_i - n_i p_i}{(1 - p_i)^2} = -\frac{n_i}{p_i} - \frac{n_i}{(1 - p_i)} = -\frac{n_i}{p_i(1 - p_i)}$$

$$\mathrm{E}\left[\frac{\partial^2 l_n(p_1, p_2)}{\partial p_0 \partial p_1}\right] = \mathrm{E}\left[\frac{\partial^2 l_n(p_1, p_2)}{\partial p_1 \partial p_0}\right] = \mathrm{E}\frac{\partial}{\partial p_0}\left[\frac{X_1}{p_1^2} - \frac{n_1 - X_1}{(1 - p_1)^2}\right] = 0$$

Значит, информационная матрица Фишера $I(p_1, p_2) = \begin{pmatrix} \frac{n_1}{p_1(1 - p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1 - p_2)} \end{pmatrix}$

c) Найдем асимптотическую стандартную ошибку, используя многопараметрический дельта-метод. Здесь $\psi = g(p_1, p_2) = p_1 - p_2$.

$\widehat{se}(\hat{\psi}) = \sqrt{(\hat{\nabla} g)^T \hat{J}_n (\hat{\nabla} g)}$, где $\hat{J}_n = J_n(\hat{p}_1, \hat{p}_2)$, $\hat{\nabla} g = \nabla g(p_1 = \hat{p}_1, p_2 = \hat{p}_2)$

$$J_n(p_1, p_2) = I_n^{-1}(p_1, p_2) = \begin{pmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{pmatrix}$$

$$\nabla g(p_1, p_2) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Тогда $\hat{se}(\hat{\psi}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

d)

In [111]:

```
from scipy import stats as st

n1 = 200
n2 = 200
X1 = 160
X2 = 148

p1 = X1/n1  # оценки вероятностей из пункта a
p2 = X2/n2
MLE = p1 - p2

se = (p1*(1-p1)/n1 + p2*(1-p2)/n2) ** .5  # оценка стандартного отклонения из пункт
z = st.norm.ppf(.05, 0, 1) # вычислим квантиль
print("ОМП для ψ: %.2f" % MLE)
print("90% " + "доверительный интервал для ψ:  [%.5f" % (MLE + z * se) + ", %.5f]"
```

```
ОМП для ψ: 0.06
90% доверительный интервал для ψ:  [-0.00904, 0.12904]
```

In [115]:

```
import numpy as np

N = 50  # количество элементов в изначальной выборке
alpha = 0.1
bin_data1 = np.random.binomial(n1, p1, N)
bin_data2 = np.random.binomial(n2, p2, N)
bin_data = np.array([[bin_data1[i], bin_data2[i]] for i in range(N)])

B = 50000  # фиксируем количество бутстреп-выборок
values = []
for i in range (B):
    boot_data = bin_data[np.random.randint(0, N, 1)]  # генерируем бутстреп-выборку
    values += [boot_data[0][0]/n1 - boot_data[0][1]/n2]
values = np.sort(np.array(values))

Xq1 = int(B * alpha/2 - 1)
Xq2 = int(B * (1 - alpha/2))
print("Эфронов 90% " + "доверительный интервал для ψ:  [%.5f" % values[Xq1] + ", %.
```

```
Эфронов 90% доверительный интервал для ψ:  [-0.00500, 0.11500]
```

Задание 3

а) Поскольку ОМП не зависит от параметризации, то ОМП для σ^2 будет равна квадрату ОМП для σ. Используя то же свойство, ОМП для τ [95% квантили $\mathcal{N}(\mu, \sigma^2)$] является 95% квантиль распределения $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, где $\hat{\mu}$ и $\hat{\sigma}$ $-$ ОМП для μ и σ соответственно. Значит, искомая оценка максимального правдоподобия $\hat{\tau} = \hat{\mu} + 1.645\hat{\sigma}$ (из таблицы нормального распределения).

Найдем $\hat{\mu}$ и $\hat{\sigma}$. Функция правдоподобия имеет вид:

$$\mathcal{L}_n(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(X_i - \mu)^2}{2\sigma^2} \right\} = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{ -\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2} \right\}$$

Логарифмическая функция правдоподобия:

$$l_n(\mu, \sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}$$

Откуда

$$\frac{\partial l_n}{\partial \mu} = 2\sum_{i=1}^{n} \frac{X_i - \mu}{2\sigma^2} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma^2}$$

$$\frac{\partial l_n}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^3}$$

$$\frac{\partial^2 l_n}{\partial \mu^2} = \frac{\partial}{\partial \mu}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma^2}\right) = \frac{n}{\sigma^2} = -\frac{n}{\sigma^2} < 0$$

$$\frac{\partial^2 l_n}{\partial \mu \partial \sigma} = \frac{\partial^2 l_n}{\partial \sigma \partial \mu} = \frac{\partial}{\partial \sigma}\left(\frac{-\sum_{i=1}^{n} X_i + n\mu}{2\sigma^2}\right) = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma^3} = [\text{при } \mu = \bar{X}] = 0$$

$$\frac{\partial^2 l_n}{\partial \sigma^2} = \frac{n}{\sigma^2} - 3\sum_{i=1}^{n}\frac{(X_i - \mu)^2}{\sigma^4} = \left[\text{при } \sigma = \sqrt{\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{n}}\right] = \frac{n}{\sigma^2} - 3\frac{n\sigma^2}{\sigma^4} = -\frac{2n}{\sigma^2} < 0$$

Значит, искомые ОМП (выражаем через выборочное среднее $\bar{X}$ и выборочную дисперсию S^2):

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n}\frac{(X_i - \mu)^2}{n}} = S$$

$$\hat{\tau} = \bar{X} + 1.645S$$

b) Найдем доверительный интервал, используя многопараметрический дельта-метод. Здесь $\tau = g(\mu, \sigma) = \mu + 1.645\sigma$.

Сначала найдем $\widehat{se}(\hat{\tau}) = \sqrt{(\hat{\nabla} g)^T \hat{J}_n (\hat{\nabla} g)}$, где $\hat{J}_n = J_n(\hat{\mu}, \hat{\sigma})$, $\hat{\nabla} g = \nabla g(\mu = \hat{\mu}, \sigma = \hat{\sigma})$.

Используя производные, посчитанные выше, получим:

$$\mathrm{E}\left[\frac{\partial^2 l_n}{\partial \mu^2}\right] = -\frac{n}{\sigma^2}$$

$$\mathrm{E}\left[\frac{\partial^2 l_n}{\partial \mu \partial \sigma}\right] = \mathrm{E}\left[\frac{\partial^2 l_n}{\partial \sigma \partial \mu}\right] = \mathrm{E}\left[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma^3}\right] = \frac{\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] - n\mu}{\sigma^3} = 0$$

$$\mathrm{E}\left[\frac{\partial^2 l_n}{\partial \sigma^2}\right] = \frac{n}{\sigma^2} - 3\mathrm{E}\left[\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^4}\right] = \frac{n}{\sigma^2} - 3\frac{n\sigma^2}{\sigma^4} = -\frac{2n}{\sigma^2}$$

Значит, информационная матрица Фишера $I_n(\mu, \sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}$

Поэтому $J_n(\mu, \sigma) = I_n^{-1}(\mu, \sigma) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix}$

$\nabla g(\mu, \sigma) = \begin{pmatrix} 1 \\ 1.645 \end{pmatrix}$

Таким образом, получим $\hat{se}(\hat{\tau}) = \sqrt{\frac{1}{n}(\hat{\sigma}^2 + \frac{\hat{\sigma}^2}{2}1.645^2)} = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1 + \frac{1.645^2}{2}}$

Тогда границы приближенного $1 - \alpha$ доверительного интервала для τ имеют вид:

$$\bar{X} + 1.645S \pm \sqrt{1 + \frac{1.645^2}{2}} z_{\alpha/2} \frac{S}{\sqrt{n}}$$

c) Используем результаты из пунктов a) и b)

In [191]:

```
import numpy as np

data = np.array([3.23, -2.50, 1.88, -0.68, 4.43, 0.17,
1.03, -0.07, -0.01, 0.76, 1.76, 3.18,
0.33, -0.31, 0.30, -0.61, 1.52, 5.43,
1.54, 2.28, 0.42, 2.33, -1.03, 4.00,
0.39])

MLE = data.mean() + 1.645 * data.std()
se = (1 + 1.645**2/2)**.5 * data.std()/(data.size**.5)

print("ОМП для τ: %.5f" % MLE)
print("Стандартная ошибка: %.5f" % se)
```

```
ОМП для τ: 4.18068
Стандартная ошибка: 0.55761
```

In [199]:

```
N = 50  # фиксируем количество элементов в изначальной выборке
norm_data =  np.random.normal(data.mean(), data.var(), N)
B = 50000  # фиксируем количество бутстреп-выборок
values = []
for i in range (B):
    boot_data = np.random.choice(norm_data, norm_data.size)  # генерируем бутстреп-
    values += [boot_data.mean() + 1.645 * boot_data.std()]
print("Стандартная ошибка, вычисленная при помощи бутстрепа: %.5f" % (np.array(valu
```

```
Стандартная ошибка, вычисленная при помощи бутстрепа: 0.61551
```

Задание 4

a) Заметим, что $\psi = P(Y_1 = 1) = P(X_1 > 0) = P(-X_1 < 0) = P(-X_1 + \theta < \theta) = \Phi(\theta)$, поскольку $X_1 \sim \mathcal{N}(\theta, 1) \Rightarrow -X_1 + \theta \sim \mathcal{N}(0, 1)$

Поскольку ОМП не зависит от параметризации, то ОМП для $\psi = \Phi(\theta)$ будет равна $\hat{\psi} = \Phi(\hat{\theta})$, где $\hat{\theta}$ − ОМП для θ.

Найдем $\hat{\theta}$. Функция правдоподобия имеет вид:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(X_i - \theta)^2}{2}\right\} = \frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2}\right\}$$

Логарифмическая функция правдоподобия:

$$l_n(\theta) = -\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2}$$

Откуда

$$\frac{\partial l_n}{\partial \theta} = 2\sum_{i=1}^{n} \frac{X_i - \theta}{2} = \sum_{i=1}^{n} X_i - n\theta$$

$$\frac{\partial^2 l_n}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left(\sum_{i=1}^{n} X_i - n\theta\right) = -n < 0$$

Значит искомые ОМП $\hat{\theta} = \overline{X}$ и $\hat{\psi} = \Phi(\overline{X})$.

b) Согласно дельта-методу, поскольку Φ является гладкой функцией, то

$$\hat{se}(\hat{\psi}) = |\Phi'(\hat{\theta})| \hat{se}(\hat{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\hat{\theta}^2}{2}\right\} \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{2\pi n e^{\left(\hat{\theta}^2\right)}}} = \frac{1}{\sqrt{2\pi n e^{\left(\overline{X}^2\right)}}}$$

Получили приближенный 95% доверительный интервал для ψ:

$$\left[\Phi\,(\bar{X}) - \frac{1.96}{\sqrt{2\pi n e^{\left(\bar{X}^2\right)}}},\ \Phi\,(\bar{X}) + \frac{1.96}{\sqrt{2\pi n e^{\left(\bar{X}^2\right)}}} \right]$$

c) По закону больших чисел $\sum_{i=1}^{n} \frac{Y_i}{n}$ является состоятельной оценкой для $\mathrm{E}Y_1$.

Поэтому $\tilde{\psi} = \sum_{i=1}^{n} \frac{Y_i}{n} \xrightarrow[n \to \infty]{\mathrm{P}} \mathrm{E}Y_1 = \mathrm{E}I\{X_1 > 0\} = P(X_1 > 0) = P(Y_1 = 1) = \psi$, а значит $\tilde{\psi}$ является состоятельной оценкой для ψ, ЧТД

d) Согласно дельта-методу, поскольку Φ является гладкой функцией, то

$$se(\hat{\psi}) = |\Phi'(\theta)| se(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\theta^2}{2}\right\} \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{2\pi n e^{\left(\theta^2\right)}}}$$

Из независимости Y_i:

$$\mathrm{V}\tilde{\psi} = \mathrm{V}\bar{Y} = n\frac{\mathrm{V}Y_i}{n^2} = \frac{\mathrm{E}Y_1^2 - (\mathrm{E}Y_1)^2}{n} = \frac{\mathrm{E}Y_1 - (\mathrm{E}Y_1)^2}{n} = [\text{по пунктам a) и c)}] = \frac{\Phi\,(\theta) - (\Phi\,(\theta))^2}{n} = \frac{\Phi\,(\theta)(1 - \Phi\,(\theta))}{n} =$$

Поэтому $se(\tilde{\psi}) = \sqrt{\frac{\Phi\,(\theta)\,\Phi\,(-\theta)}{n}}$

Таким образом:

$$\sqrt{n}(\tilde{\psi} - \psi) \rightsquigarrow \mathcal{N}(0,\ \Phi\,(\theta)\,\Phi\,(-\theta))$$

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow \mathcal{N}\left(0, \frac{1}{2\pi e^{\left(\theta^2\right)}}\right)$$

Значит, искомая относительная асимптотическая эффективность $\tilde{\psi}$ к $\hat{\psi}$:

$$ARE(\tilde{\psi}, \hat{\psi}) = \frac{1}{2\pi e^{\left(\theta^2\right)}\,\Phi\,(\theta)\,\Phi\,(-\theta)}$$

e) Аналогично пункту c) по закону больших чисел и из непрерывности функции Φ получим:

$$\hat{\psi} = \Phi\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) \xrightarrow[n \to \infty]{\mathrm{P}} \Phi\,(\mathrm{E}X_1) = P(\mathrm{E}X_1 \geq Z) = P(\mathrm{E}X_1 - Z > 0) \neq P(X_1 > 0) = \psi$$

где $Z \sim \mathcal{N}(0, 1)$. Значит, по определению оценка не является состоятельной, ЧТД.

Неравенство (в редких случаях все-таки может быть и равенство) следует из предположения, что данные на самом деле не распределены нормально, поэтому распределение X_1 отличается от распределения $\mathrm{E}X_1 - Z \sim \mathcal{N}(\mathrm{E}X_1, 1)$.

В то же время есть сходимость по вероятности при $n \to \infty$ к величине $\Phi\,(\mathrm{E}X_1)$.

Задание 5

a) Функция мощности

$$W(\theta) = P_\theta(Y > c) = 1 - P_\theta(X_{(n)} \le c) = 1 - P_\theta(X_1 \le c, X_2 \le c, \ldots, X_n \le c) = 1 - \prod_{i=1}^{n} P_\theta(X_i \le c) = \begin{cases} 1 - \prod_{i=1}^{n} 0 = 1, \\ 1 - \prod_{i=1}^{n} \dfrac{c}{\theta} = 1 - \dfrac{c^n}{\theta^n}, \\ 1 - \prod_{i=1}^{n} 1 = 0, \end{cases}$$

b) По определению размер критерия

$$\alpha = \sup_{\theta \in \Theta_0} W(\theta)$$

Поскольку здесь $\Theta_0 = \left\{\frac{1}{2}\right\}$, то получим условие на искомый параметр c (сразу учтем, что нужная мощность отлична от 0 и 1):

$$0.05 = 1 - (2c)^n \Rightarrow \text{ c } = \frac{0.95^{\frac{1}{n}}}{2}$$

c) В данной задаче критерий размера α, построенный для статистики $Y = X_{(n)}$, имеет вид: H_0 отвергается, если $Y > c_\alpha$.

Поэтому по теореме из лекций $p\text{-}value = P_{\theta_0}(X_{(n)} > 0.48) =$ [в данном случае $0 < c = 0.48 < 0.5 = \theta$] $= 1 - (2 \cdot 0.48)^{20} \approx 0.558$

Так как $p\text{-}value > 0.1$, то ничего определенного о гипотезе H_0 сказать нельзя.

In [126]:

```
print("p-value: %.5f" % (1-(2 * 0.48)**20))
```

```
p-value: 0.55800
```

d) Аналогично предыдущему пункту здесь $p\text{-}value = P_{\theta_0}(X_{(n)} > 0.52) =$ [в данном случае $c = 0.52 > 0.5 = \theta$] $= 0$. Поэтому гипотеза H_0 заведома неверна.

Задание 6

a) Для каждого лекарства рассмотрим выборку $X_1, \ldots, X_n$, где n количество людей, принимавших данное лекарство. Будем считать, что X_j принимает значение 0, если у j-ого пациента случилось осложнение, иначе $-$ 0. Тогда X_i $-$ независимые одинаково распределенные случайные величины из распределения Бернулли с параметром p.

Найдем ОМП для p.

Функция распределения имеет вид: $f(x, p) = p^x(1-p)^{1-x}$, где $x = 0, 1$

Функция правдоподобия:

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i, p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{n\overline{X}}(1-p)^{n-n\overline{X}}$$

Логарифм функции правдоподобия:

$$l_n(p) = n\overline{X}\ln(p) + (n - n\overline{X})\ln(1-p)$$

$$\frac{\partial l_n(p)}{\partial p} = \frac{n\overline{X}}{p} - \frac{n - n\overline{X}}{1-p} = n\frac{\overline{X} - p}{p(1-p)}$$

$$\frac{\partial^2 l_n(p)}{\partial p^2} = -\frac{n\overline{X}}{p^2} - \frac{n - n\overline{X}}{(1-p)^2} < 0$$

Последнее неравенство следует из того, что $\overline{X} < 1$ (т.к. $X_i <= 1\ \forall i$)

Откуда искомая оценка $\hat{p} = \overline{X}$

Обозначим вероятность "успешности" i-го лекарства p_i (в порядке их следования в задании), а "успешности" i-го по сравнению с Placebo $pl_i = p_i - p_0$

Поскольку ОМП не зависит от параметризации, то ОМП для $pl_i = p_i - p_0$ будет равна $\widehat{pl_i} = \hat{p}_i - \hat{p}_0$, где $\hat{p}_i$ и $\hat{p}_0$ $-$ оценки максимального правдоподобия для p_i и p_0 соответственно.

Найдем, исходя из этого, оценки "успешности" лекарств ($\hat{p}_i$) и оценки их "успешности" относительно Placebo ($\widehat{pl_i}$).

In [342]:

```
import pandas as pd

data = pd.DataFrame()
data["Лекарство"] = ["Placebo", "Chlorpromazine", "Dimenhydrinate", "Pentobarbital
data["Кол-во пациентов"] = [80, 75, 85, 67, 85]
data["Кол-во осложнений"] = [45, 26, 52, 35, 37]
data["Оценка успешности"] = (data["Кол-во пациентов"] - data["Кол-во осложнений"])/
data["Оценка успешности отн. Placebo"] = data["Оценка успешности"] - data["Оценка у
data
```

Out[342]:

	Лекарство	Кол-во пациентов	Кол-во осложнений	Оценка успешности	Оценка успешности отн. Placebo
0	Placebo	80	45	0.437500	0.000000
1	Chlorpromazine	75	26	0.653333	0.215833
2	Dimenhydrinate	85	52	0.388235	-0.049265
3	Pentobarbital (100 mg)	67	35	0.477612	0.040112
4	Pentobarbital (150 mg)	85	37	0.564706	0.127206

Протестируем "успешность" каждого из лекарств по сравнению с Placebo (кроме него самого) на 5% уровне значимости. Для этого проверим гипотезы $H_{0i}: pl_i = 0$ против $H_{1i}: pl_i \neq 0$ по критерию Вальда для $i > 0$.

Найдем стандартные ошибки pl_i

$\mathrm{V}\hat{p}_k =$ [из независимости X_i (элементов выборки для k-го лекарства)] =
$\frac{1}{n_k^2}\sum_{i=1}^{n_k} \mathrm{V}X_i = \frac{1}{n_k^2}\sum_{i=1}^{n_k} p_k(1-p_k) = \frac{1}{n_k^2} n p_k(1-p_k) = \frac{p_k(1-p_k)}{n_k}$

$\mathrm{V}pl_i =$ [из независимости, т.к. рассматриваем отличные от 0 индексы] $= \mathrm{V}p_i + \mathrm{V}p_0 = \frac{p_i(1-p_i)}{n_i} + \frac{p_0(1-p_0)}{n_0}$

Откуда $\hat{se}(pl_i) = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_i} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}}$

Теперь можно построить критерий Вальда размера α.

$$|W_i| = \left|\frac{\hat{pl}_i - 0}{\hat{se}(pl_i)}\right| = \left|\frac{\hat{pl}_i}{\hat{se}(pl_i)}\right|$$

Гипотеза H_{0i} отклоняется, если $|W_i| > z_{\alpha/2}$

In [343]:

```
from scipy.stats import norm
from math import *

z = -norm.ppf(.025, 0, 1)  # найдем квантиль

data["Станд. ошибка"] = data["Оценка успешности"]
for i in range(5):  # вычислим оценки согласно формулам выше
    data.loc[i, "Станд. ошибка"] = \
    (data.loc[i, "Оценка успешности"]*(1 - data.loc[i, "Оценка успешности"])/ data.
     data.loc[0, "Оценка успешности"]*(1 - data.loc[0, "Оценка успешности"])/data.l

data.drop([0], inplace=True)
data["Гипотеза принята (Вальд)"] = (abs(data["Оценка успешности отн. Placebo"]/data
data["p-value"] = 2*norm.cdf(-abs(data["Оценка успешности отн. Placebo"]/data["Стан
data
```

Out[343]:

	Лекарство	Кол-во пациентов	Кол-во осложнений	Оценка успешности	Оценка успешности отн. Placebo	Станд. ошибка	Гипотеза принята (Вальд)	p
1	Chlorpromazine	75	26	0.653333	0.215833	0.078077	False	0.0
2	Dimenhydrinate	85	52	0.388235	-0.049265	0.076618	True	0.5
3	Pentobarbital (100 mg)	67	35	0.477612	0.040112	0.082462	True	0.6
4	Pentobarbital (150 mg)	85	37	0.564706	0.127206	0.077253	True	0.0

b) Проверим гипотезы по методам Бонферрони и Benjamini-Hochberg.

По теореме из лекций для критерия Вальда $p\text{-}value \simeq P(|Z| > |W|) = 2\,\Phi\,(-|W|)$, где $Z \sim \mathcal{N}(0, 1)$.

In [346]:

```
alpha = 0.05

data["Гипотеза принята (Бонферрони)"] = (data["p-value"] > alpha/data.shape[0])

pvalues = np.sort(np.array([data["p-value"][i] for i in range(1, data.shape[0] + 1)
cm = 1  # т.к. p-values независимы (поскольку независимы лекарства)
val_ii = pvalues - np.array([i*alpha/(cm*data.shape[0]) for i in range (1, data.sha
T = pvalues[np.argmax(val_ii < 0)]  # пороговое значение метода Benjamini-Hochberg
data["Гипотеза принята (Benjamini-Hochberg)"] = (data["p-value"] > T)

data
```

Out[346]:

	Лекарство	Кол-во пациентов	Кол-во осложнений	Оценка успешности	Оценка успешности отн. Placebo	Станд. ошибка	Гипотеза принята (Вальд)	p
1	Chlorpromazine	75	26	0.653333	0.215833	0.078077	False	0.0
2	Dimenhydrinate	85	52	0.388235	-0.049265	0.076618	True	0.5
3	Pentobarbital (100 mg)	67	35	0.477612	0.040112	0.082462	True	0.6
4	Pentobarbital (150 mg)	85	37	0.564706	0.127206	0.077253	True	0.0

P.S. Можно также посмотреть на знак оценки успешности относительно Placebo для отвергнутых гипотез, откуда видно, что из всех лекарств только Chlorpromazine успешнее Placebo на заданном уровне значимости (для всех тестов).

Задание 7

а) Сначала найдем оценку максимального правдоподобия для параметра Пуассоновского распределения.

Функция распределения имеет вид: $f(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$

Функция правдоподобия:

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} f(X_i, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} = e^{-n\lambda} \frac{\lambda^{n\bar{X}}}{\prod_{i=1}^{n} X_i!}$$

Логарифм функции правдоподобия:

$$\mathrm{l}_n(\lambda) = -n\lambda + n\bar{X}\ln(\lambda) - \ln(\prod_{i=1}^{n} X_i!)$$

$$\frac{\partial l_n(\lambda)}{\partial \lambda} = -n + \frac{n\bar{X}}{\lambda}$$

$$\frac{\partial^2 l_n(\lambda)}{\partial \lambda^2} = -\frac{n\bar{X}}{\lambda^2}$$

Единственным корнем $\frac{\partial l_n(\lambda)}{\partial \lambda} = 0$ является $\lambda = \bar{X}$, и поскольку в этой точке вторая производная отрицательна (т.к. $\lambda > 0$), то оценка $\hat{\lambda} = \bar{X}$ является искомой ОМП. По теореме из лекций оценка ОМП является асимптотически нормальной.

Информация Фишера:

$$I_n(\lambda) = -\mathrm{E}\frac{\partial^2 l_n(\lambda)}{\partial \lambda^2} = -\mathrm{E}\left[-\frac{n\bar{X}}{\lambda^2}\right] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

Откуда $\hat{se} = \frac{1}{\sqrt{I_n(\hat{\lambda})}} = \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{\frac{\bar{X}}{n}}$

Теперь можно построить критерий Вальда размера α.

$$|W| = \left|\frac{\hat{\lambda} - \lambda_0}{\hat{se}}\right| = \left|\sqrt{\frac{n}{\bar{X}}}(\bar{X} - \lambda_0)\right|$$

Искомый критерий: гипотеза H_0 отклоняется, если $\left|\sqrt{\frac{n}{\bar{X}}}(\bar{X} - \lambda_0)\right| > z_{\alpha/2}$

b)

In [57]:

```
import numpy as np
from scipy import stats as st
from math import *

lam0 = 1
n = 20
alpha = 0.05
z = -st.norm.ppf(alpha/2, 0, 1) # вычислим квантиль

pois_data =  np.random.poisson(lam0, n)
W = abs((n/pois_data.mean())**.5 * (pois_data.mean() - lam0))
print("W: %.5f, " % W +"z: %.5f" % z)
if (W > z):
    print("Гипотеза H_0 отклоняется")
else:
    print("Гипотеза H_0 принимается")
```

```
W: 1.34715, z: 1.95996
Гипотеза H_0 принимается
```

In [55]:

```
acc = 0
den = 0
num = 10000
for i in range(num):
    pois_data =  np.random.poisson(lam0, n)
    W = abs((n/pois_data.mean())**.5 * (pois_data.mean() - lam0))
    if W <= z:
        acc += 1
    else:
        den += 1
print("Принята: " + str(acc) + " раз, отклонена: " + str(den) + " раз.")
print("Доля ошибок 1 рода: %.5f" % (den/num))
```

```
Принята: 9477 раз, отклонена: 523 раз.
Доля ошибок 1 рода: 0.05230
```

Доля ошибок первого рода получилась очень близка к $\alpha = 0.05$.

Задание 8

Найдем оценку максимального правдоподобия для p.

Распределение вероятности имеет вид: $f(x, p) = C_n^x p^x (1-p)^{n-x}$

Функция правдоподобия: $\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i, p) = \prod_{i=1}^{n} C_n^{X_i} p^{X_i} (1-p)^{n-X_i}$

Логарифм функции правдоподобия: $l_n(p) = \sum_{i=1}^{n} \left[\ln(C_n^{X_i}) + X_i \ln(p) + (n - X_i)\ln(1-p) \right]$

Найдем производные:

$$\frac{\partial l_n(p)}{\partial p} = \sum_{i=1}^{n}\left[\frac{X_i}{p} - \frac{n - X_i}{1-p}\right] = \sum_{i=1}^{n}\left[\frac{X_i - X_ip - np + X_ip}{p(1-p)}\right] = \sum_{i=1}^{n}\left[\frac{X_i - np}{p(1-p)}\right] = \frac{n\bar{X} - n^2p}{p(1-p)}$$

$$\frac{\partial^2 l_n(p_1, p_2)}{\partial p_i^2} = \sum_{i=1}^{n}\left[-\frac{X_i}{p^2} - \frac{n - X_i}{(1-p)^2}\right] < 0$$

Последнее неравенство следует из того, что $0 \le X_i \le n$.

Поэтому искомая ОМП $\hat{p} = \dfrac{\bar{X}}{n}$

Статистика отношения правдоподобий принимает вид (выразим через $\hat{p}$, используя, что $\bar{X} = n\hat{p}$):

$$\lambda = 2\ln\frac{\mathcal{L}_n(\hat{p})}{\mathcal{L}_n(\hat{p}_0)} = 2\ln\frac{\prod_{i=1}^{n} C_n^{X_i}\hat{p}^{X_i}(1-\hat{p})^{n-X_i}}{\prod_{i=1}^{n} C_n^{X_i}p_0^{X_i}(1-p_0)^{n-X_i}} =$$

$$= 2\ln\prod_{i=1}^{n}\left(\frac{\hat{p}}{p_0}\right)^{X_i}\left(\frac{1-\hat{p}}{1-p_0}\right)^{n-X_i} = 2\ln\left[\left(\frac{\hat{p}}{p_0}\right)^{n^2\hat{p}}\left(\frac{1-\hat{p}}{1-p_0}\right)^{n^2-n^2\hat{p}}\right] =$$

$$= 2n^2\left(\hat{p}\ln\left(\frac{\hat{p}}{p_0}\right) + (1-\hat{p})\ln\left(\frac{1-\hat{p}}{1-p_0}\right)\right)$$

Тогда по теореме из лекций для данного критерия p-$value \simeq P(\chi_1^2 > \lambda) = P(Z^2 > \lambda) = P(|Z| > \sqrt{\lambda}) = 2\,\Phi(-\sqrt{\lambda})$, где $Z \sim \mathcal{N}(0, 1)$.

Найдем теперь статистику Вальда.

Информация Фишера:

$$I_n(p) = -\mathrm{E}\sum_{i=1}^{n}\left(-\frac{X_i}{p^2} - \frac{n - X_i}{(1-p)^2}\right) = \mathrm{E}\left[\frac{n\bar{X}}{p^2} + \frac{n^2 - n\bar{X}}{(1-p)^2}\right] = \frac{n^2p}{p^2} + \frac{n^2 - n^2p}{(1-p)^2} = \frac{n^2}{p} + \frac{n^2}{1-p} = \frac{n^2}{p(1-p)}$$

Откуда $\hat{se} = \dfrac{1}{\sqrt{I_n(\hat{p})}} = \dfrac{\sqrt{\hat{p}(1-\hat{p})}}{n}$

Значит, статистика Вальда $W = \dfrac{\hat{p} - p_0}{\hat{se}} = \dfrac{n(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})}}$

По теореме из лекций для данного критерия p-$value \simeq P(|Z| > |W|) = 2\,\Phi(-|W|)$, где $Z \sim \mathcal{N}(0, 1)$.

Сравним полученные тесты аналитически, сравнив p-$value$. Для этого посмотрим насколько похожи $\sqrt{\lambda}$ и W или, что то же самое, $\lambda = n^2\left(2\hat{p}\ln\left(\frac{\hat{p}}{p_0}\right) + 2(1-\hat{p})\ln\left(\frac{1-\hat{p}}{1-p_0}\right)\right)$ и $W^2 = n^2\frac{(\hat{p}-p_0)^2}{\hat{p}(1-\hat{p})}$.

Поскольку $\hat{p}$ − ОМП для p, то она является состоятельной оценкой для p, и поэтому $\hat{p} \xrightarrow[n\to\infty]{\mathrm{P}} p$. А значит, при больших n: $W^2 \sim \lambda \sim n^2$.

Таким образом, построенные тесты очень похожи.

Теперь сравним их экспериментально.

In [172]:

```
import numpy as np
from math import *
from scipy.stats import norm

N = 200  # количество элементов в выборке
p = 0.8
bin_data = np.random.binomial(N, p, N)
p_est = bin_data.mean()/N
p_0 = 0.804

p_val_l = 2 * norm.cdf(-(2 * N**2 * (p_est * log(p_est/p_0) + (1-p_est)* log((1- p_
p_val_w = 2 * norm.cdf(- abs(N * (p_est - p_0)/(p_est * (1 - p_est))** .5))  # для
print("p_value критерия отношения правдоподобий: %.5f" % p_val_l + ", p_value крите
```

```
p_value критерия отношения правдоподобий: 0.13141, p_value критерия Ва
льда: 0.13288
```

In [190]:

```
diff = 0
num = 10000  # столько раз будем генерировать данные
N = 200  # количество элементов в выборке
p = 0.8
p_var = np.random.normal(0, 0.005, num)  # построим p_0 с помощью нормального шума
pn = np.array([p for _ in range (num)])
pn_0 = pn + p_var
for i in range (num):
    bin_data = np.random.binomial(N, p, N)
    p_est = bin_data.mean()/N
    p_0 = pn_0[i]
    p_val_l = 2 * norm.cdf(-(2 * N**2 * (p_est * log(p_est/p_0) + (1-p_est)* log((1
    p_val_w = 2 * norm.cdf(- abs(N * (p_est - p_0)/(p_est * (1 - p_est))** .5))
    diff += abs(p_val_l - p_val_w)
print("Средний модуль разности p-values данных тестов: %.5f" % (diff/num))
```

```
Средний модуль разности p-values данных тестов: 0.00061
```

Экспериментально также убедились в том, что построенные критерии работают почти совсем одинаково.