Смирнов - Контрольная 2

Problem 1

We use the Kolmogorov's criterion for homogenicity. Treat 12 months as discrete probability space. The empiric distribution function is then:

```
import numpy as np
total_children = 88273
children_per_month = np.array([7280, 6957, 7883, 7884, 7892, 7609, 7585, 7393, 7203, 6903, 6552, 7132])
empirical_cmf =  np.array([sum(children_per_month[0:i]) for i in range(1,12)])/total_children
print(empirical_cmf)
```

`output: [0.08247142 0.16128374 0.25058625 0.33990008 0.42930454 0.51550304 0.60142966 0.6851812 0.76678033 0.84498091 0.91920519]`

The Kolmogorov-Smirnov statistics is

$$\sqrt{n}D_n = \sqrt{n}\max_x |\hat{F}_n(x) - F(x)|$$

where $\hat{F}_n$ is the empirical cumulative mass function, and $F(x)$ is the theoretical cumulative mass function. In our case, the theoretical $F(x)$ is just the $\frac{1}{12}\lceil x \rceil$ function, so

```
empirical_cmf = np.array([0.08247142, 0.16128374, 0.25058625, 0.33990008, 0.42930454, 0.51550304, 0.60142966, 0.6851812, 0.76678033, 0.84498091, 0.91920519])
theoretical_cmf = (1/12)*np.array([x for x in range(1, 12)])
print(theoretical_cmf)
Kolmogorov_Smirnov = np.max(np.abs(empirical_cmf-theoretical_cmf))
print(Kolmogorov_Smirnov)
print(f"Criterion:{((12)**(1/2))*Kolmogorov_Smirnov}")
```

```
output:
[0.08333333 0.16666667 0.25 0.33333333 0.41666667 0.5 0.58333333 0.66666667 0.75 0.83333333 0.91666667]
0.018514533333333416
Criterion: 0.06413622482352208
```

From the table of quantiles of Kolmogorov's distribution

$$K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}$$

we obtain for $\alpha = 0.05$ $Z_{1-\alpha} = 1.36$. Hence, for $\alpha = 0.05$ the homogenicity hypothesis is not rejected.

Problem 2

Consider a sample of size n from the Bernoulli distribution with parameter p and hypotheses $p = \frac{1}{2}, p = \frac{3}{4}$, criterions $\delta_1 : p = \frac{1}{2} \iff \bar{X} \leq \frac{1}{2}$, $\delta_2 : p = \frac{1}{2} \iff \bar{X} < 1/3$. Limits of probabilities of errors of the first and the second kind for these criteria are:

Criterion δ_1

$$\lim_{n\to\infty} P\left(\delta_1(X_n) : p = \frac{1}{2} \middle| p = \frac{3}{4}\right) = 0$$

a.s. by the Big Numbers Law: after infinite number of tries the mean will be in the $\epsilon-$neighborhood of $\frac{3}{4}$ a.s..

$$\lim_{n\to\infty} P\left(\delta_1(X_n) : p = \frac{3}{4} \middle| p = \frac{1}{2}\right) = 0.5$$

By symmetricity of the binomial/normal distribution the liimit probability of rejecting the true $p = \frac{1}{2}$ hypothesis is 0.5.

Criterion δ_2

$$\lim_{n\to\infty} P\left(\delta_2(X_n) : p = \frac{1}{2} \middle| p = \frac{3}{4}\right) = 0$$

again by the Law of Big Numbers and the same logic as with δ_1: the mean will be $> \frac{1}{3}$ a.s. As for the errors of the second kind,

$$\lim_{n\to\infty} P\left(\delta_2(X_n) : p = \frac{3}{4} \middle| p = \frac{1}{2}\right) = 1$$

by exactly the same reason.

Problem 3

The contigency table is:

```
cont_table = [[15729000, 3497],
                          [16799000, 3072]]
```

Let us use the Chi-squared test with the null hypothesis that gender and deaf-muteness are indpendent, and the number of people with death-muteness in a population is a Gaussian random variable. Then the value of the Chi-squared statistics is

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed number of people with deaf-muteness in each gender, and E_i is the expected value, which we take for

$$E_i = \frac{\text{Total number of people with deaf-muteness}}{\text{Total number of people}} \cdot \text{Number of people of gender } i$$

```
total = 15729000 + 16799000
total_deaf_mute = 3497 + 3072
proportion = total_deaf_mute/total
E_male = 15729000*proportion
E_female = 16799000*proportion
O_male = 3497
O_female = 3072

print(proportion, E_male, E_female)
```

```
chi_sq = (O_male - E_male)**2/E_male + (O_female - E_female)**2/E_female
print(f"Chi-squared statistics: {chi_sq}")
```

```
output:

0.0002019490900147565 3176.457236842105 3392.5427631578946
Chi-squared statistics: 62.6329505033744
```

The number of parameters of the χ^2 distribution is 1, because there are two classes. Hence, for $\alpha = 0.05$ the critical value of the χ^2 statistics is 3.841, so the null hypothesis is rejected and the deaf-mutedness is dependent on the geneder.