МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования

«Дальневосточный федеральный университет»
(ДВФУ)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Департамент математического и компьютерного моделирования

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА И СЛУЧАЙНЫЕ ПРОЦЕССЫ

ЛАБОРАТОРНАЯ РАБОТА 3

Тема: Эмпирическая функция распределения

Студент
Гузовская Александра Чеславовна
группы Б9123-01.03.02сп

Преподаватель Деревягин А. А.

Регистрационный № ___________ Оценка________________________

___________ _________________ ___________ _________________
(подпись) (И. О. Фамилия) (подпись) (И. О. Фамилия)

«___» __________________ 2025 г. «___» __________________ 2025 г.

г. Владивосток
2025

Лабораторная работа 3
Эмпирическая функция распределения

Гузовская Александра Чеславовна
Б9123-01.03.02сп

05 апреля 2025

1 Матчасть. Эмпирическая и теоретическая функции распределения

Теоретическая функция распределения описывает распределение вероятностей для случайной величины, основана на предположениях о её распределении. Это нормальное, биномиальное, пуассоновское или другое распределение, которое описывается математически.

Эмпирической функцией распределения, кратко ЭФР
(empirical distribution function, или функция распределения выборки)
называют функцию $\hat{F}_X(x)$, определяющую для каждого значения $x_i, i = 0, 1, .., n$ относительную частоту события X < x.

Получается по определению

$$\hat{F}_n(x) = \frac{n_x}{n}$$

где
n_x - число вариант, меньших x,
n – объем выборки.

В отличие от эмпирической функции распределения выборки, функцию распределения генеральной совокупности называют теоретической

функцией распределения.
Различия между ними таковы, что
- теоретическая функция определяет вероятность события X < x,
- эмпирическая – относительную частоту этого события.

Другой вид эмпирической функции распределения, идентичный первому

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

где I — индикаторная функция.

2 Матчасть. Связь с истинной функцией распределения

Истинная функция распределения $F(x)$ — это функция, которая описывает вероятность того, что случайная величина X примет значение, меньшее или равное x. Она является теоретической и основана на распределении, из которого была получена выборка.

При увеличении размера выборки n (числа наблюдений)
ЭФР $\hat{F}_X(x)$ будет все более точно приближаться к истинной функции распределения $F(x)$.
ЭФР стремится к истинной функции распределения $F(x)$ при увеличении размера выборки n.

Теорема:
ЭФР $\hat{F}_X(x)$ сходится к истинной функции распределения $F(x)$ почти наверное (то есть с вероятностью 1)

Теорема Гливенко-Кантелли:

$$\sup_x \left|\hat{F}_n(x) - F_X(x)\right| \xrightarrow{P} 0$$

является более строгой, чем центральная предельная теорема

3 Матчасть. Доверительный интервал

Доверительный интервал — это диапазон значений, в котором с заданной вероятностью находится истинное значение параметра.
в свою очередь 95%-й доверительный интервал это доверительный интервал, который охватывает 95% возможных значений истинного параметра

Доверительный интервал для стандартного отклонения можно вычислить с использованием хи-квадрат распределения:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}\right)$$

где
s — выборочное стандартное отклонение,
χ^2 — критические значения хи-квадрат распределения.

Для эмпирической функции распределения в материалах практики указана формула для ε для нахождения доверительного интервала:

$$\varepsilon_n = \sqrt{\frac{1}{2n} \cdot ln(\frac{2}{\alpha})}$$

где α - уровень значимости, $1-\alpha$ - уровень доверия, а нижняя и верхняя границы интервала определяются следующим образом:

$$U(x) = min(\hat{F}_n(x) + \varepsilon_n, 1);$$

$$L(x) = max(\hat{F}_n(x) - \varepsilon_n, 0);$$

4 Матчасть. Центральная предельная теорема или ЦПТ

Теорема утверждает, что когда случайная величина формируется путем сложения большого числа независимых случайных величин, дисперсии которых малы по сравнению с дисперсией суммы, распределение этой величины приближается к нормальному.

Чем больше независимых слагаемых в сумме, тем ближе распределение к нормальному. Вместо суммы рассматривается и среднее арифметическое случайных величин, отличается от суммы множителем $\frac{1}{n}$, его распределение также стремится к нормальному при увеличении числа n.

В случае с ЭФР используется для того, чтоб доказать её сходимость к ИФР

5 Ограничения использования ЭФР и ЦПТ

- При малых выборках ЭФР может значительно колебаться и не точно отражать истинное распределение. ЦПТ также требует, чтобы размер выборки был достаточно большим

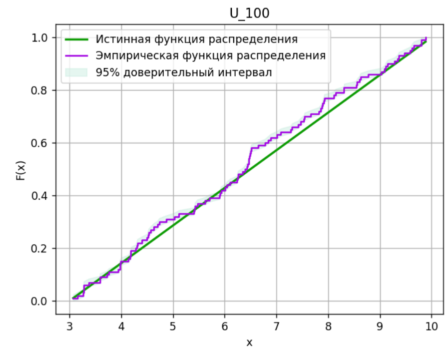
- Если данные зависимы или имеют разные распределения, ЭФР может не корректно оценивать истинное распределение и ЦПТ может не применяться.

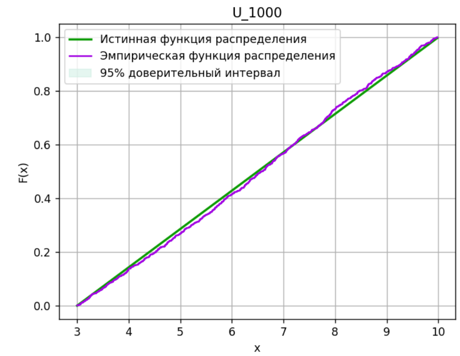
- ЭФР может не точно отражать поведение распределения на границах (в крайних значениях)

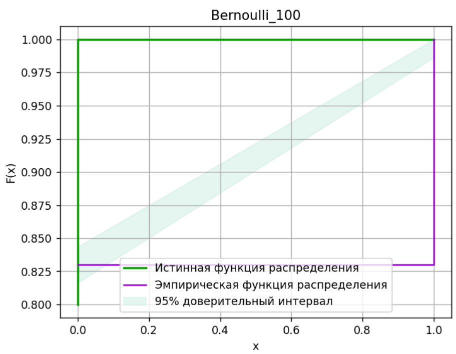
- Если величины имеют разные распределения, то ЦПТ может не сработать.

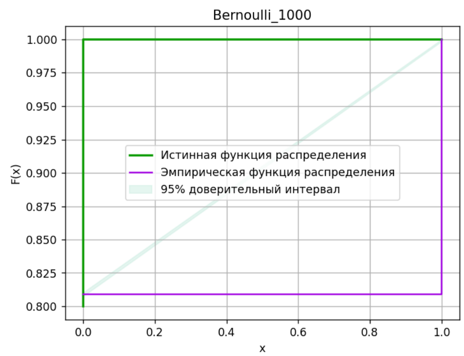
- Если распределение имеет бесконечную дисперсию, ЦПТ не применима.

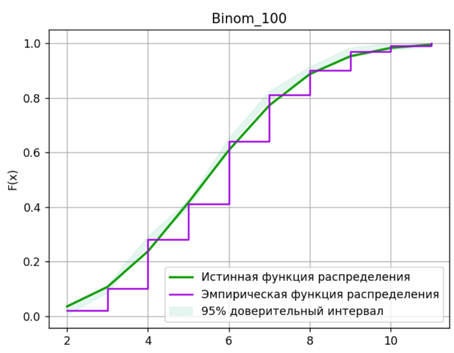
6 Графики

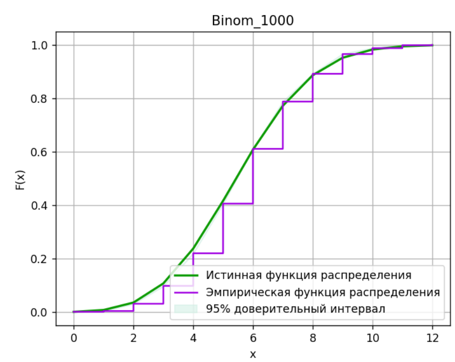

U_100
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x

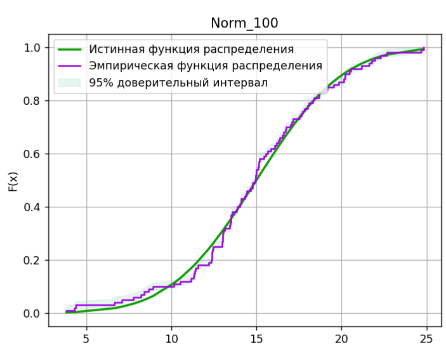

U_1000
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x

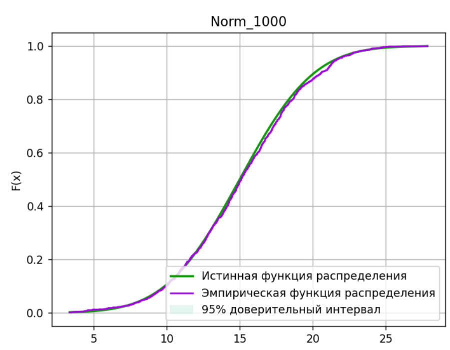

Bernoulli_100
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x


Bernoulli_1000
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x


Binom_100
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x


Binom_1000
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x


Norm_100
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x


Norm_1000
Истинная функция распределения
Эмпирическая функция распределения
95% доверительный интервал
F(x)
x

7 Код программы

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

random_state = 9

samples_ = {
    'U_100': stats.uniform.rvs(loc=3, scale=7, size=100,
    random_state=random_state),
    'U_1000': stats.uniform.rvs(loc=3, scale=7, size=1000,
    random_state=random_state),
    'Bernoulli_100': stats.bernoulli.rvs(p=0.2, size=100,
    random_state=random_state),
    'Bernoulli_1000': stats.bernoulli.rvs(p=0.2, size=1000,
    random_state=random_state),
    'Binom_100': stats.binom.rvs(n=20, p=0.3, size=100,
    random_state=random_state),
    'Binom_1000': stats.binom.rvs(n=20, p=0.3, size=1000,
    random_state=random_state),
    'Norm_100': stats.norm.rvs(loc=15, scale=4, size=100,
    random_state=random_state),
    'Norm_1000': stats.norm.rvs(loc=15, scale=4, size=1000,
    random_state=random_state),
}

def empiricalDistributionFunction(data, x, alpha=0.05):
    """
    Вычисляет эмпирическую функцию распределения
    и 95% доверительный интервал.

    :param data: массив данных
    :param x: значение, для которого вычисляется ЭФР
    :param alpha: уровень значимости
    :return: значение ЭФР и доверительный интервал
    """
    n = len(data)
```

```
    edf_value = np.sum(data <= x) / n
    error_ = np.sqrt((1/(2*n) * np.log(2/alpha)) / n)

    bottom_ = max(0, edf_value - error_)
    top_ = min(1, edf_value + error_)

    return edf_value, (bottom_, top_)

def makePlot(label, data):
    """
    Создает график для эмпирической функции распределения
    и истинной функции распределения.
    Учитывает, что для дискретных распределений
    Бернулли и Биномиального
    не применяется построение через plot

    :param label: название распределения
    :param data: массив данных
    """
    x_values = np.sort(data)
    true_cdf = None

    if 'U_' in label:
        true_cdf = stats.uniform.cdf(x_values, loc=3, scale=7)
    elif 'Bernoulli' in label:
        true_cdf = stats.bernoulli.cdf(np.arange(0, 2), p=0.2)
        true_cdf = np.array([stats.bernoulli.cdf(x, p=0.2) for x in x_values])
    elif 'Binom' in label:
        true_cdf = np.array([stats.binom.cdf(k, n=20, p=0.3) for k in x_values])
    elif 'Norm' in label:
        true_cdf = stats.norm.cdf(x_values, loc=15, scale=4)

    empirical_cdf = [empiricalDistributionFunction(data, x, label)[0] for x in x_v

    confidence_intervals = [empiricalDistributionFunction(data, x, label)[1]
    for x in x_values]
    bottoms_ = [ci[0] for ci in confidence_intervals]
    tops_ = [ci[1] for ci in confidence_intervals]
```

```
    plt.figure()

    plot_params = {
        'label': ['Истинная функция распределения',
        'Эмпирическая функция распределения', '95% доверительный интервал'],
        'color': ['#0e9c05', '#a304e3', '#cdeee3'],
        'linewidth': [2, 1.5, 1]
    }

    if ("Bernoulli" or "Binom_") in label:
        plt.step(x_values, true_cdf, label=plot_params['label'][0],
        color=plot_params['color'][0], linewidth=plot_params['linewidth'][0])
    else:
        plt.plot(x_values, true_cdf, label=plot_params['label'][0],
        color=plot_params['color'][0], linewidth=plot_params['linewidth'][0])

    plt.step(x_values, empirical_cdf, label=plot_params['label'][1],
    color=plot_params['color'][1], linewidth=plot_params['linewidth'][1],
    where='post')
    plt.fill_between(x_values, bottoms_, tops_, label=plot_params['label'][2],
    color=plot_params['color'][2], linewidth=plot_params['linewidth'][2],
    alpha=0.5)

    plt.title(label)
    plt.xlabel('x')
    plt.ylabel('F(x)')
    plt.legend()
    plt.grid()
    plt.show()

for label, data in samples_.items():
    makePlot(label, data)
```

https://github.com/huzouskaya/math_statistics