Кафедра дискретной математики МФТИ

Курс математической статистики

Игашов Илья, 593 групппа

Задание №2

№4.

Сгенерируйте выборку $X_1, \ldots, X_N$ из стандартного нормального распределения для $N = 10^4$. Для всех $n \in N$ посчитайте по ней эмпирическую функцию распределения. Для некоторых n (например, $n \in \{10, 25, 50, 100, 1000, N\}$) постройте графики эмпирической функции распределения (отметьте на оси абсцисс точки "скачков" кривых, нанеся каждую из "подвыборок" на ось абсцисс на каждом соответствующем графике с коэффициентом прозрачности 0.2), нанеся на каждый из них истинную функцию распределения (количество графиков равно количеству различныз значений n). Для всех $n \in N$ посчитайте точное значение $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ и постройте график зависимости статистик D_n и $\sqrt{n} D_n$ от n.

In [1]:

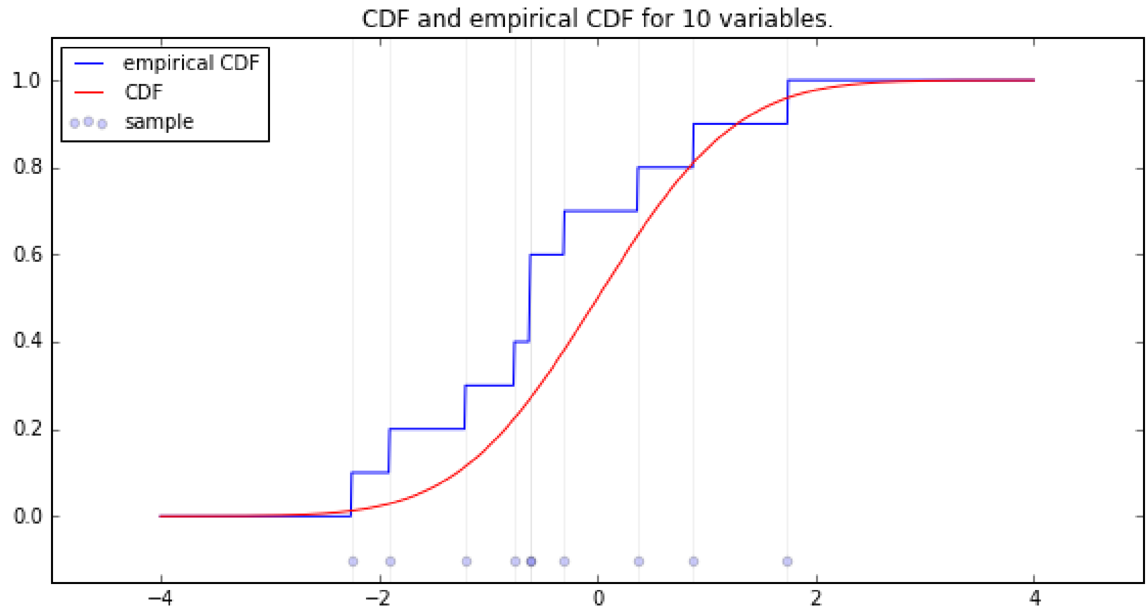
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
from statsmodels.distributions.empirical_distribution import ECDF

%matplotlib inline
```

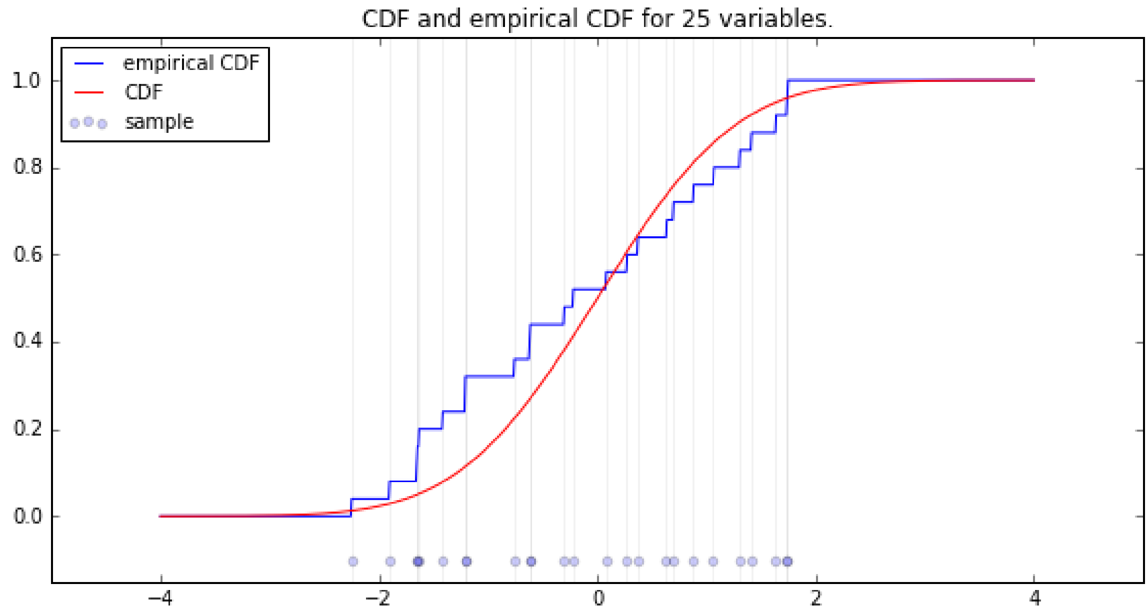
In [2]:

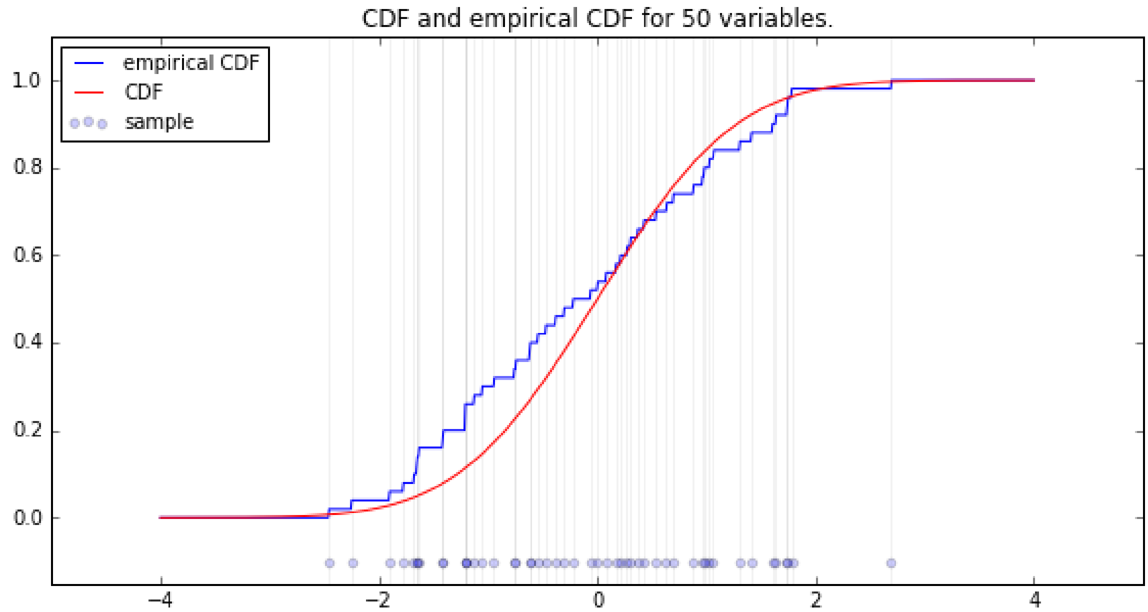
```
# Генерируем выборку размера N = 10000 из стандартного нормального распределения.
N = 10000
sample = norm.rvs(size=N)
```

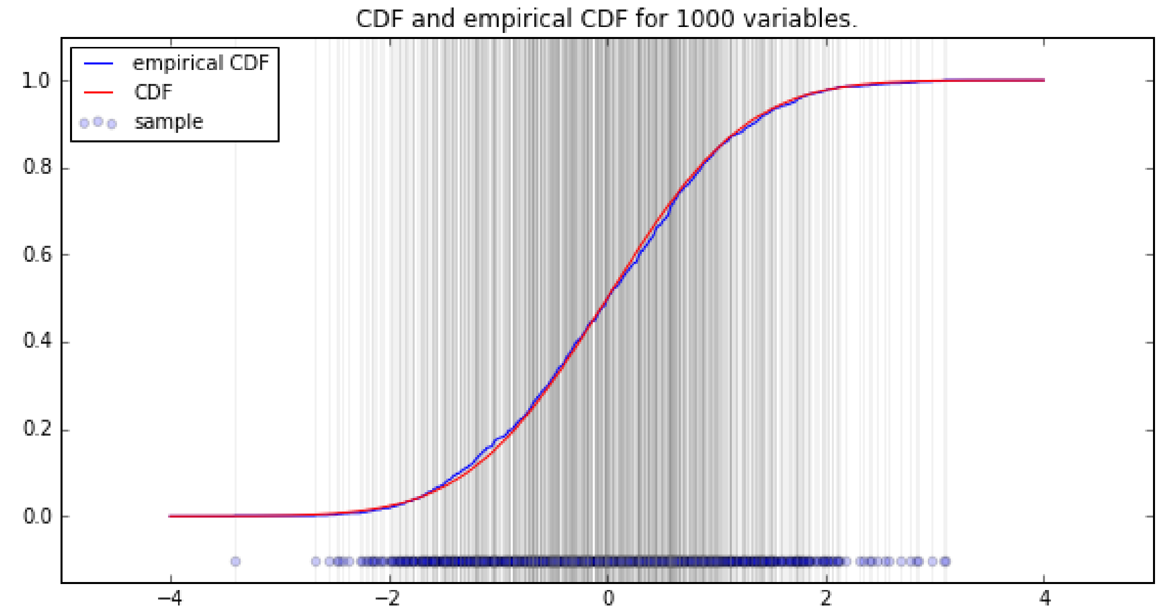
In [3]:

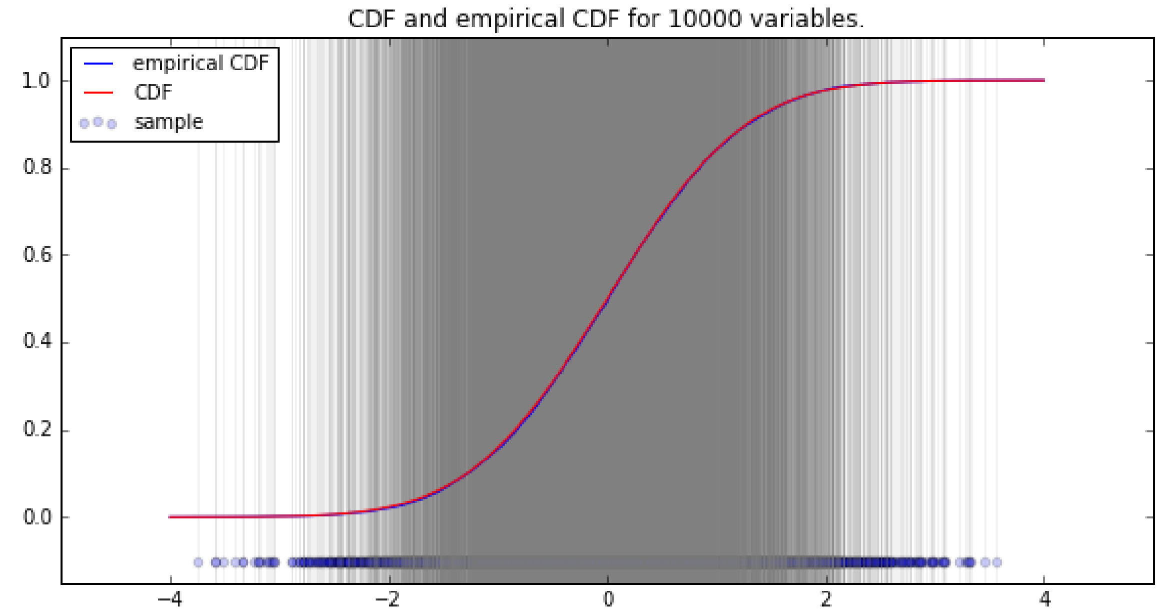
```
# Для n = 10, 25, 50, 1000, 10000 строим графики эмпирической ф.р., нанося на каждый и
з них также истинную ф.р.
for n in [10, 25, 50, 1000, N]:
    grid = np.linspace(-4, 4, 1000)  # Задаем сетку для построения графика ф.р.
    ecdf = ECDF(sample[:n])  # Эмпирическая ф.р.
    emp_cdf = ecdf(grid)  #  Набор значений эмпирической ф.р. на наборе аргументов gri
d.

    plt.figure(figsize=(10, 5))
    plt.plot(grid, emp_cdf, label='empirical CDF')
    plt.scatter(sample[:n], np.zeros(n) - 0.1, alpha=0.2, label='sample')
    plt.plot(grid, norm.cdf(grid), color='red', label='CDF')
    # Для обозначения скачков воспользуемся вертикальными линиями погрешностей.
    plt.errorbar(sample[:n], np.zeros(n) - 1, yerr=2*n, alpha=0.1, barsabove=Tru
e, color='grey')

    plt.ylim(-0.15, 1.1)
    plt.xlim(-5, 5)
    plt.title(r'CDF and empirical CDF for ' + str(n) + ' variables.')
    plt.legend(fontsize=10, loc=2)
    plt.show()
```


CDF and empirical CDF for 10 variables.
empirical CDF
CDF
sample
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4


CDF and empirical CDF for 25 variables.
empirical CDF
CDF
sample
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4


CDF and empirical CDF for 50 variables.
empirical CDF
CDF
sample
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4





Для всех $n \in N$ посчитаем значение $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ и построим график зависимости статистик D_n и $\sqrt{n} D_n$ от n.

Высчитывать значения $F(x)$ и $F_n(x)$ будем только в точках скачков эмпирической ф.р., то есть в точках $X_1, \ldots, X_n$. При этом, поскольку мы ищем точную грань, для каждой точки X_i будем вычислять $F_n(X_i)$ и $F_n(X_i - \epsilon)$, а ϵ положим равным 0.0001. Будем сравнивать величины $|F_n(X_i) - F(X_i)|,\ |F_n(X_i - \epsilon) - F(X_i)|$ и выбирать большую, а потом среди полученных значений выберем максимальное.

In [4]:

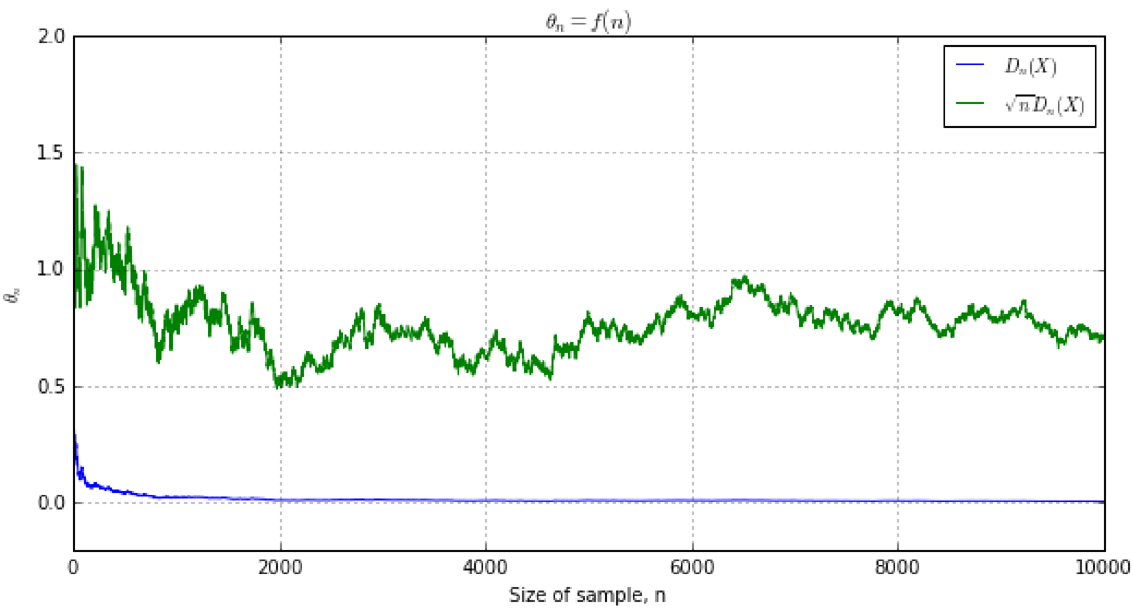
```
N = 10000
norm_cdf = norm.cdf(sample)  # Значения функции распределения на элементах выборки.

Dn = np.array([np.max(np.absolute(ECDF(sample[:n])(sample[:n]) - norm_cdf[:n]))
            for n in range(1, N + 1, 1)]) # Статистика Dn: вычисляем значения ecdf
 точно в точках выборки.

eps = 0.0001
Dneps = np.array([np.max(np.absolute(ECDF(sample[:n])(sample[:n] - eps) - norm_c
df[:n]))
            for n in range(1, N + 1, 1)]) # Статистика Dneps: вычисляем значения ec
df в точках Xi - eps.
```

In [5]:

```
# Статистики. Каждое значение статистики Dn (назовем ее variance) - это максимум из зна
чений Dn и Dneps для каждого n.
variance = np.array([max(Dn[n], Dneps[n]) for n in range(N)])
n_variance = np.array([variance[n] * ((n + 1) ** 0.5) for n in range(N)])
```

In [6]:

```
# Графики:
plt.figure(figsize=(10, 5))
plt.plot(np.linspace(1, N, N), variance, label='$D_n(X)$')
plt.plot(np.linspace(1, N, N), n_variance, label='$\sqrt{n}D_n(X)$')
plt.ylim(-0.2, 2)

plt.title(r'$\theta_n = f(n)$')
plt.xlabel(r'Size of sample, n', fontsize='10')
plt.ylabel(r'$\theta_n$', fontsize='10')
plt.grid()
plt.legend(fontsize=10, loc=1)
plt.show()
```



Вывод

Мы рассмотрели эмпирическую функцию стандартного нормального распределения, построили ее график для разных размеров n выборки и, построив график зависимости статистики $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ от n, убедились в справедливости теоремы Гливенко-Кантелли:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \ \text{ п. н. } \ \text{ при } \ n \to \infty.$$