Введение в анализ данных

Домашнее задание 3.

Правила:

- Дедлайн **17 мая 23:59**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru` , указав тему письма `"[номер группы] Фамилия Имя - Задание 3"` . Квадратные скобки обязательны.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `3.N.ipynb` и `3.N.pdf` , где `N` -- ваш номер из таблицы с оценками. *pdf-версию можно сделать с помощью Ctrl+P. Пожалуйста, посмотрите ее полностью перед отправкой. Если что-то существенное не напечатается в pdf, то баллы могут быть снижены.*
- Решения, размещенные на каких-либо интернет-ресурсах, не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Для выполнения задания используйте этот ноутбук в качестве основы, ничего не удаляя из него.
- Если код будет не понятен проверяющему, оценка может быть снижена.
- Никакой код при проверке запускаться не будет.

Баллы за задание:

Легкая часть (достаточно на "хор"):

- Задача 1 -- 3 балла

Сложная часть (необходимо на "отл"):

- Задача 2 -- 2 балла
- Задача 3 -- 10 баллов
- Задача 4 -- 4 балла

Баллы за разные части суммируются отдельно, нормируются впоследствии также отдельно. Иначе говоря, 1 балл за легкую часть может быть не равен 1 баллу за сложную часть.

Легкая часть

Перед выполнением этой части настоятельно рекомендуется посмотреть ноутбук с лекции про закон больших чисел.

Задача 1.

В этой задаче нужно визуализировать *центральную предельную теорему*.

а). Пусть $\xi_1, \ldots, \xi_n$ --- независимые случайные величины из распределения $Exp(\lambda)$. Согласно центральной предельной теореме выполнена сходимость

$$Z_n = \frac{X_n - \mathsf{E}X_n}{\sqrt{\mathsf{D}X_n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

где $X_n = \sum_{i=1}^{n} \xi_i$. Вам нужно убедиться в этом, сгенерировав множество наборов случайных величин и посчитав по каждому из наборов величину Z_n в зависимости от размера набора.

Сгенерируйте 500 наборов случайных величин $\xi_1^j, \ldots, \xi_{300}^j$ из распределения $Exp(1)$. По каждому из них посчитайте сумму $X_{jn} = \sum_{i=1}^{n} \xi_i^j$ для $1 \leqslant n \leqslant 300$, то есть сумма первых n величин j-го набора. Для этого среднего посчитайте величину

$$Z_{jn} = \frac{X_{jn} - \mathsf{E}X_{jn}}{\sqrt{\mathsf{D}X_{jn}}}.$$

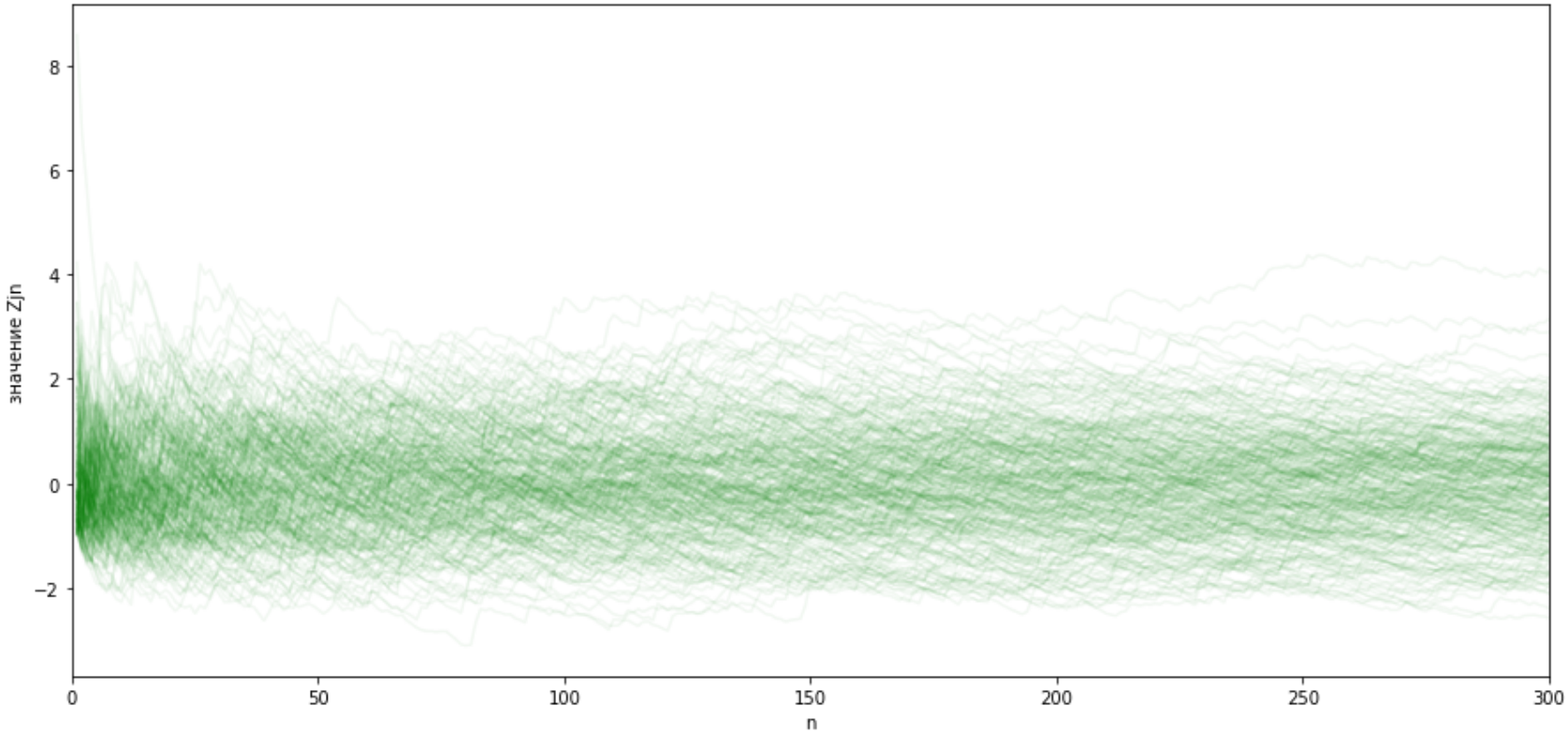
In [1]:
```
import numpy as np
import scipy.stats as sps
import matplotlib.pyplot as plt
```

In [2]:
```
size = 300
sets_amt = 500
samples = (sps.expon(1).rvs(size=(sets_amt, size)) - sps.expon(1).expect())\
                                                   / np.sqrt(sps.expon(1).var())
Z = samples.cumsum(axis=1) / np.sqrt(np.arange(1, size + 1))
```

Для каждого j нанесите на один график зависимость Z_{jn} от n. Каждая кривая должна быть нарисована *одним цветом* с прозрачностью `alpha=0.05`. Сходятся ли значения Z_{jn} к какой-либо константе?

In [3]:
```
plt.figure(figsize=(15, 7))
for i in range(size):
 plt.plot(np.arange(size) + 1, Z[i], color='green', alpha=0.05)
plt.xlabel('n')
plt.ylabel('значение Zjn')
plt.xlim((0, size))
```

Out[3]: (0, 300)

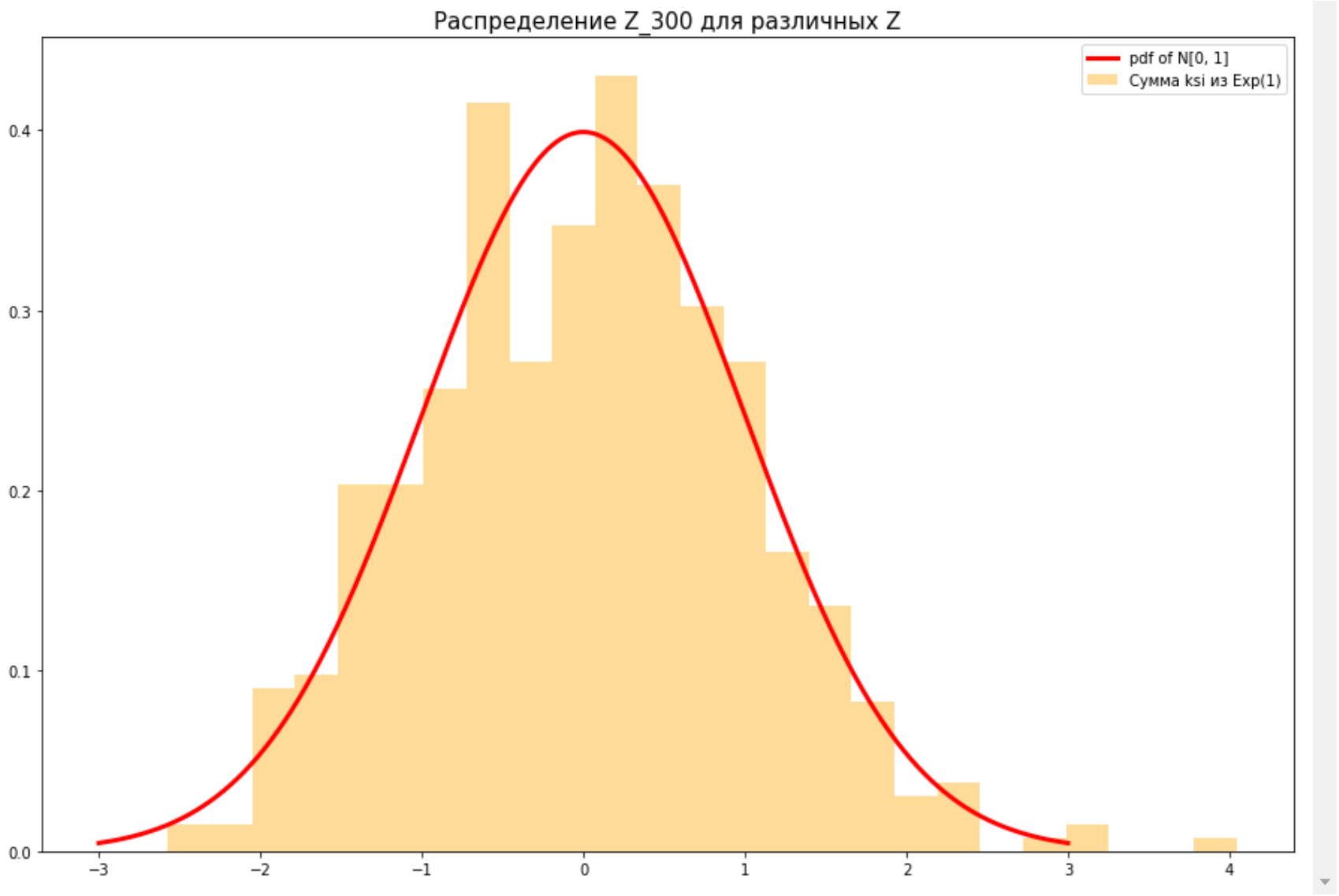


Type *Markdown* and LaTeX: α^2

Для $n = 300$ по набору случайных величин $Z_{1,300}, \ldots, Z_{500,300}$ постройте гистограмму. Похожа ли она на плотность распределения $\mathcal{N}(0, 1)$ (ее тоже постройте на том же графике)? Не забудьте сделать легенду.

In [44]:
```
grid = np.linspace(-3, 3, 1000)
plt.figure(figsize=(15, 10))
plt.hist(
    [z[-1] for z in Z],
    bins = 25,
    density = True,
    alpha = 0.4,
    color = 'orange',
    label = 'Сумма ksi из Exp(1)'
)
plt.plot(
    grid,
    sps.norm.pdf(grid),
    color = 'red',
    lw = 3,
    label = 'pdf of N[0, 1]'
)
plt.title("Распределение Z_300 для различных Z", fontsize=15)
plt.legend()
```

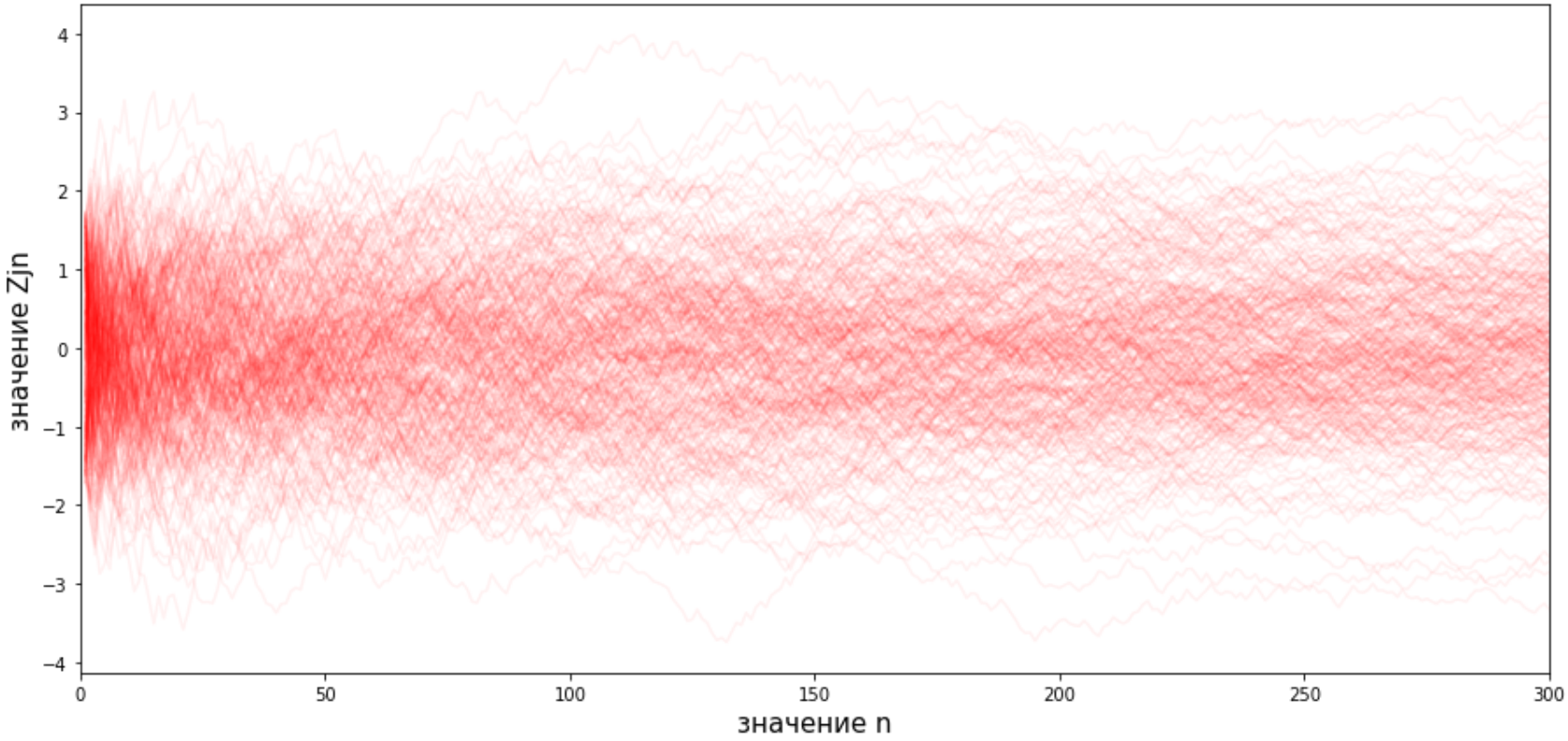
Out[44]:
```
<matplotlib.legend.Legend at 0x7fe4f7b6d4f0>
```



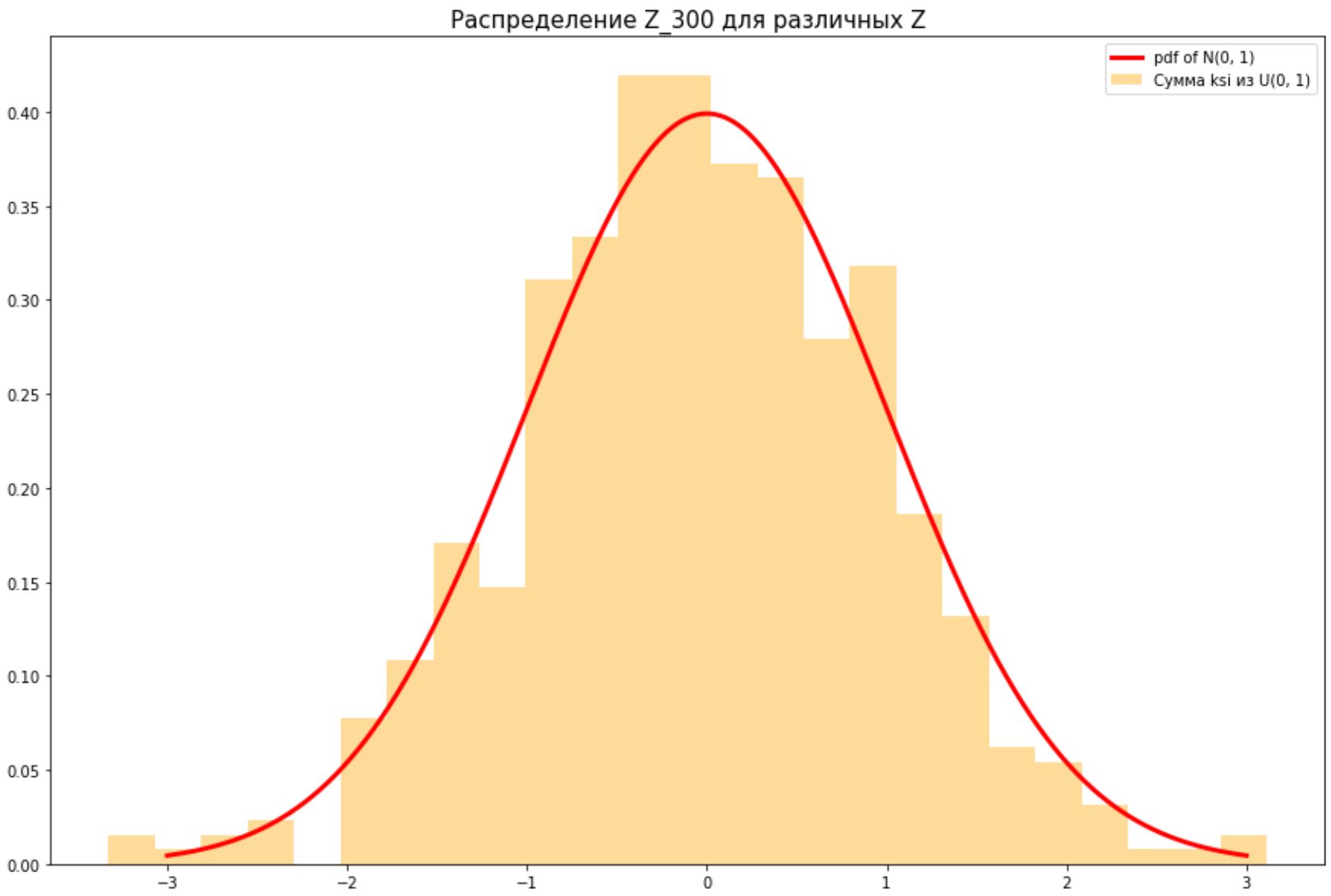
b). Выполните те же действия для распределений $U(0,1)$ и $Pois(1)$.

In [50]:
```
def create_Z(distr):
    size = 300
    sets_amt = 500
    samples = (distr.rvs(size=(sets_amt, size)) - distr.expect()) / np.sqrt(distr.var())
    Z = samples.cumsum(axis=1) / np.sqrt(np.arange(1, size + 1))
    return Z

def build_graph(Z):
    plt.figure(figsize=(15, 7))
    for i in range(size):
        plt.plot(np.arange(size) + 1, Z[i], color='red', alpha=0.05)
    plt.xlabel('значение n', fontsize=15)
    plt.ylabel('значение Zjn', fontsize=15)
    plt.xlim((0, size))

def build_hist(Z, from_=None):
    plt.figure(figsize=(15, 10))
    grid = np.linspace(-3, 3, 1000)
    plt.hist(
        [z[-1] for z in Z],
        bins = 25,
        density = True,
        alpha = 0.4,
        color = 'orange',
        label = 'Сумма ksi из ' + from_
    )
    plt.plot(
        grid,
        sps.norm.pdf(grid),
        color = 'red',
        lw = 3,
        label = 'pdf of N(0, 1)'
    )
    plt.title("Распределение Z_300 для различных Z", fontsize=15)
    plt.legend()
```

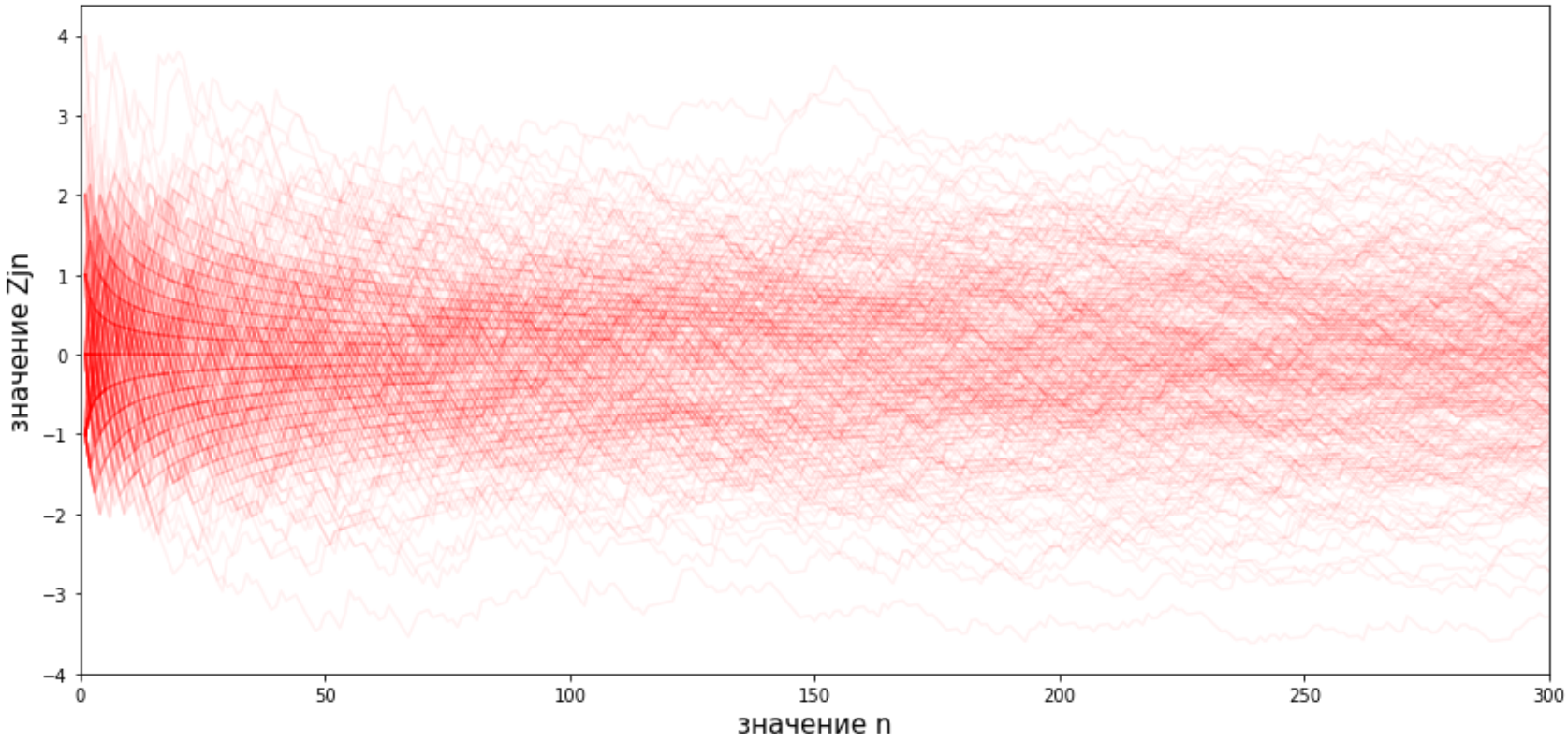
In [46]:
```
Z1 = create_Z(sps.uniform(0, 1))
build_graph(Z1)
```



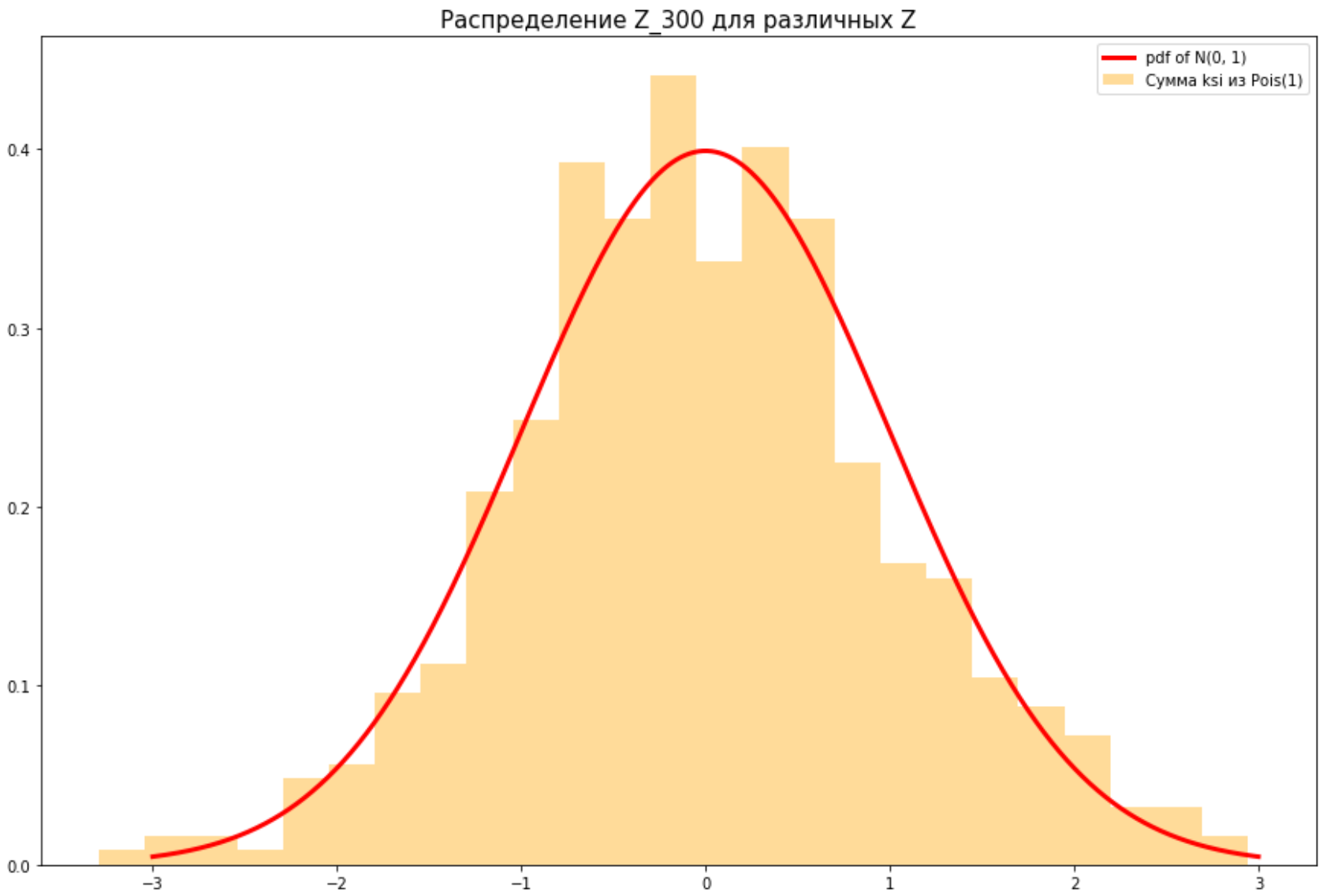
In [55]:
```
build_hist(Z1, 'U(0, 1)')
```



In [52]:
```
Z2 = create_Z(sps.poisson(1))
build_graph(Z2)
```



In [54]:
```
build_hist(Z2, 'Pois(1)')
```



Сделайте вывод о смысле центральной предельной теоремы. Подтверждают ли сделанные эксперименты теоретические свойства?

Сумма одинаково распределенных случайных величин при увеличении их количества стремится к нормальному. Экспериментально получили такой же результат.

Type *Markdown* and LaTeX: α^2

Сложная часть

Задача 2.

В этой задаче нужно визуализировать *закон повторого логарифма*.

а). Пусть $\xi_1, \ldots, \xi_n$ --- независимые случайные величины из равномерного распределения на $\{-1, 1\}$. Согласно закону повторного логарифма траектория суммы $S_n = \xi_1 + \ldots + \xi_n$ при увеличении n с вероятностью 1 бесконечное число раз пересекает границу $\pm(1-\varepsilon)\sqrt{2n\log\log n}, \varepsilon > 0$, и лишь конечное число раз пересекает границу $\pm(1+\varepsilon)\sqrt{2n\log\log n}, \varepsilon > 0$. Вам нужно убедиться в этом, сгенерировав множество наборов случайных величин и посчитав по каждому из наборов сумму в зависимости от размера набора.

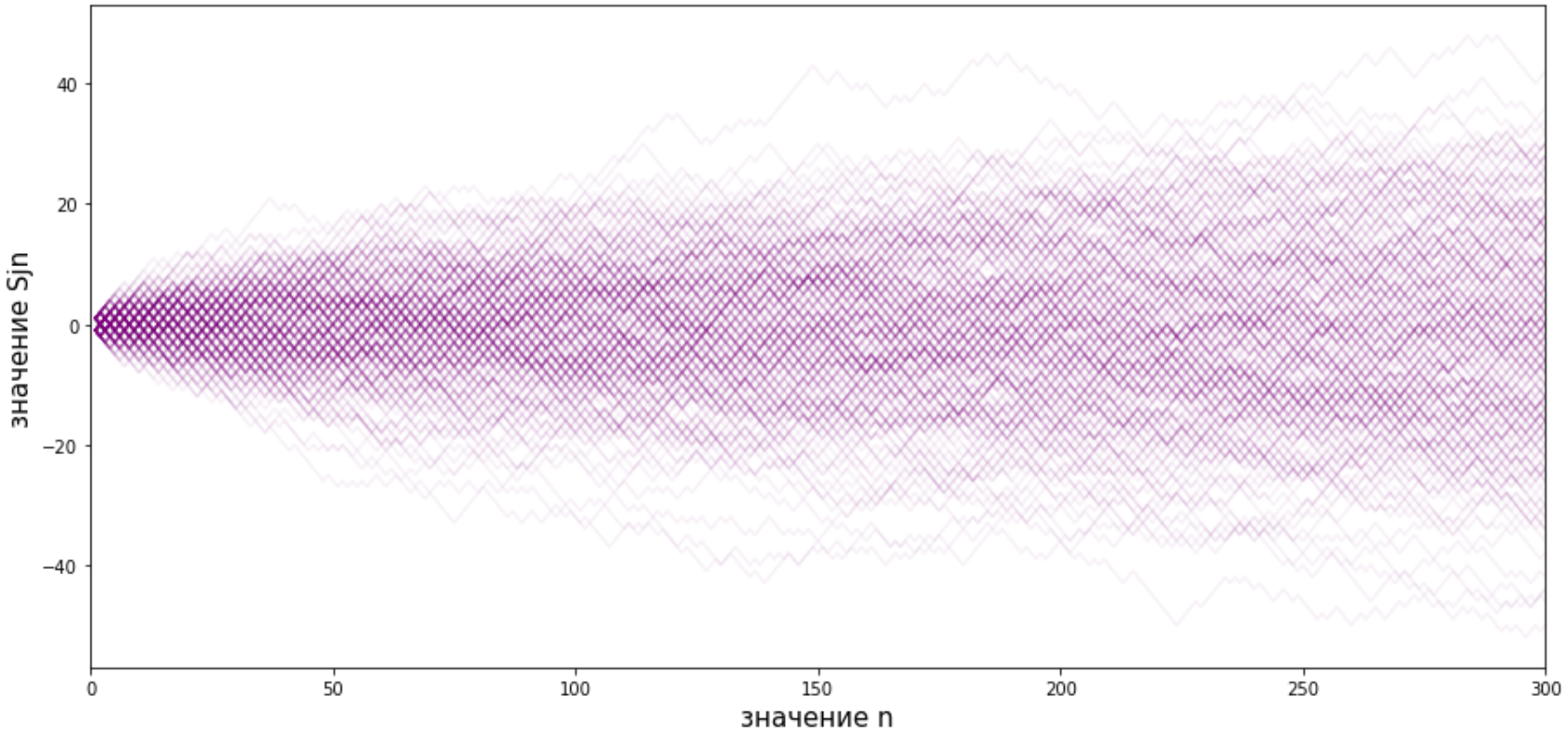
Сгенерируйте 500 наборов случайных величин $\xi_1^j, \ldots, \xi_{300}^j$ из распределения $Bern(1/2)$. По каждому из них посчитайте среднее $S_{jn} = \sum_{i=1}^{n} \xi_i^j$ для $1 \leqslant n \leqslant 300$, то есть сумму по первым n величинам j-го набора.

In [73]:
```
samples = ((sps.bernoulli(0.5).rvs(size=(sets_amt, size)) - 0.5) * 2).cumsum(axis=1)
```

Для каждого j нанесите на один график зависимость S_{jn} от n. Каждая кривая должна быть нарисована *одним цветом* с прозрачностью `alpha=0.05` .

In [74]:
```
plt.figure(figsize=(15, 7))
for i in range(size):
    plt.plot(np.arange(size) + 1, samples[i], color='purple', alpha=0.05)
plt.xlabel('значение n', fontsize=15)
plt.ylabel('значение Sjn', fontsize=15)
plt.xlim((0, size))
```

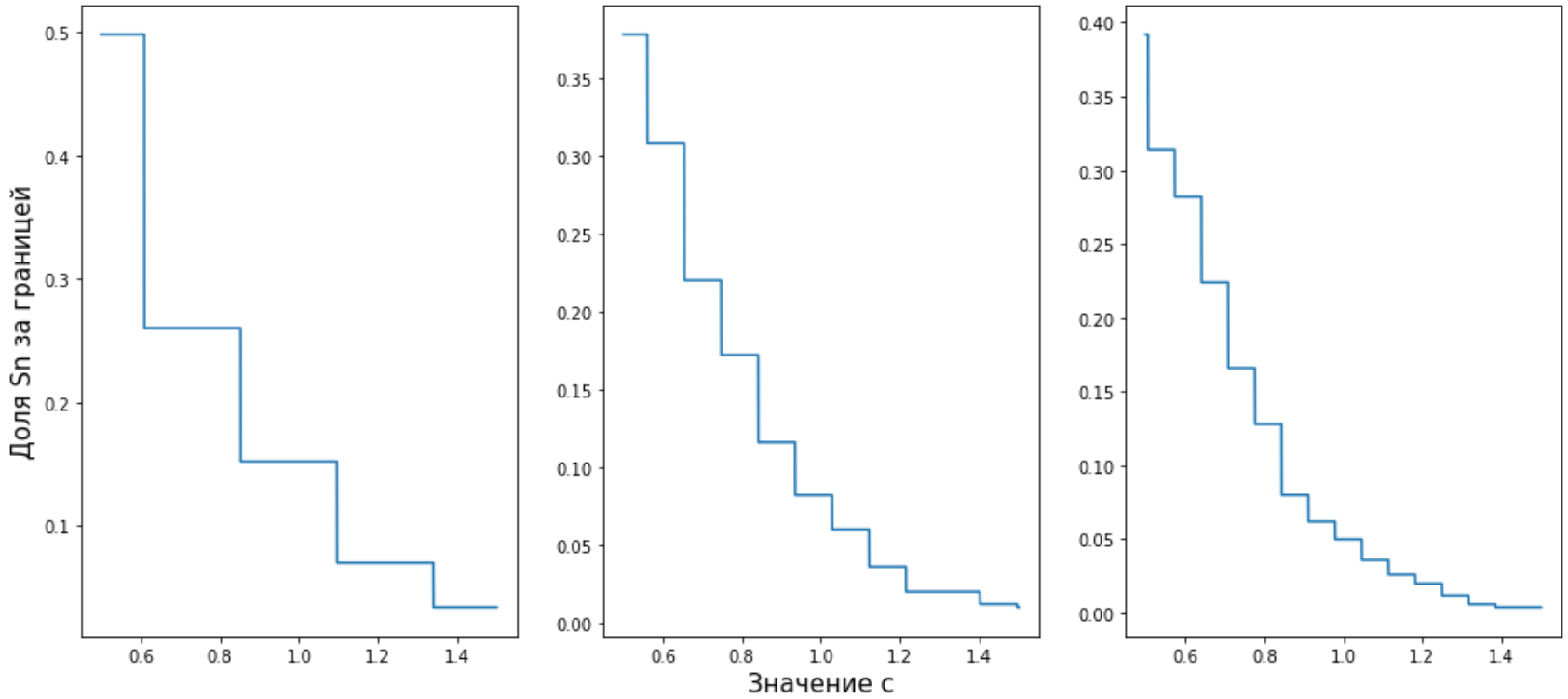
Out[74]: (0, 300)



Для трех различных n по набору случайных величин $S_{1,n}, \ldots, S_{500,n}$ постройте графих доли тех величин, которые выходят за порог $\pm c\sqrt{2n \log\log n}$ при изменении c от 0.5 до 1.5. Графики стройте в строчку с помощью `plt.subplot` .

In [75]:
```
def get_loglog(c, n):
    return  c * np.sqrt(2 * n * np.log(np.log(n)))

def get_fraction(S, border):
    return 1 - sum(1 for x in S if -border <= x <= border) / len(S)
```

In [76]:
```
plt.figure(figsize=(16,7))
n = [28, 143, 256]
c = np.linspace(0.5, 1.5, 10000)
fraction_data = np.empty(10000)
for k, i in enumerate(n):
    plt.subplot(1, 3, k + 1)
    for j, cj in enumerate(c):
        fraction_data[j] = get_fraction(samples.T[i], get_loglog(cj, i))
    plt.plot(c, fraction_data)
    if (k == 1):
        plt.xlabel('Значение c', fontsize=15)
    if (k == 0):
        plt.ylabel('Доля Sn за границей', fontsize=15)
```



Сделайте вывод о смысле закона повторного логарифма. Подтверждают ли сделанные эксперименты теоретические свойства?

Закон повторного логарифма дает информацию об отклонении случайного блуждания. Так на графиках видно, что при росте

константы, отвечающей за величину границы, вероятность пересечения этой границы стремится к нулю. То есть, наш эксперимент согласовался с теорией

Задача 3.

В этой задаче нужно проявить и визуализировать свое *творчество*.

Общий принцип:

- Придумать какую-либо цель исследования, поставить вопрос или гипотезы
- Собрать необходимый набор данных "руками" или с помощью кода.
- Сделать простой анализ полученного датасета в этом ноутбуке.
- Сделать вывод.

Основные требования к данным:

- Все собранные данные необходимо представить в виде одной или нескольких таблиц формата `csv` или `xls`. Эти файлы должны легко считываться при помощи `pandas`. **Все эти файлы необходимо прислать вместе с решением на почту.**
- По строкам таблиц должны располагаться исследуемые объекты, например, люди. Одному объекту соответствует одна строка. По столбцам должны располагаться свойства объекта, например, пол, возраст.
- При сборе данных "руками" вы самостоятельно выбираете количество исследуемых объектов исходя из времени, которое необходимо на это потратить. Рассчитываемое время -- 2-3 часа.
- При сборе данных с помощью кода ограничивайте себя только размером доступных данных, которые можно скачать за 2-3 часа или 10000 объектами.
- Во всех случаях количество исследуемых объектов должно быть **не менее 30**. Количество свойств объектов -- **не менее двух**.

Основные требования к исследованию:

- Заранее необходимо четко определиться с вопросом, который вы хотите исследовать. Например, "хочу исследовать взаимосвязь двух свойств".
- При анализе необходимо провести полную визуализацию данных. Все графики должны быть оформлены грамотно.
- Подумайте, как вы можете применить полученные математические знания по курсу теории вероятностей для анализа собранных данных?
- Примените их если это возможно. Например, у вас не должно возникнуть проблем с тем, чтобы посчитать среднее, подкрепив корректность такого подхода соответствующей теоремой. А взаимосвязь двух свойств вы вряд ли сейчас

сможете оценить по данным.
- Полноценные выводы.

Ниже перечислены некоторые идеи, но вы можете придумать свою.

- Исследование характеристик и вкусовых качеств овощей/фруктов/ягод. В качестве свойств можно рассмотреть высоту объекта, радиус в разрезе, цвет, тип, вкусовую оценку, дату покупки, дату употребления.
- Исследование характеристик листьев деревьев. В качестве свойств можно рассмотреть длину и ширину листа, цвет, тип растения.
- Характеристики товаров в интернет-магазине, включая рейтинг.
- Музыкальные исполнители и песни. В качестве свойств можно рассмотреть рейтинг артиста, количество треков, количество ремиксов, количество коллабораций.
- Кинофильмы, мультфильмы, аниме.
- Анализ новостных лент. На сайте https://www.similarweb.com/ (https://www.similarweb.com/) можно посмотреть статистику различных издательств, на основе чего придумать правило оценки степени "доверия" изданию. Исследуйте, какие новости первым публикует издание с наибольшим значением доверия? Опросите знакомых об отношении к этим новостям.
- Анализ данных пабликов ВК.
- Анализ схожести сайтов или блогов по частоте упоминания какой-либо темы.

Задача 4.

Некоторые студенты второго курса ФИВТ понадеявшись на отмену учета посещения занятий по курсу "Введение в анализ данных" решили дудосить гугл-опросники. Команда "Физтех.Статистики" без особых проблем смогла разделить результаты опроса на спамовые и настоящие, а также установить круг подозреваемых. Теперь это предлагается сделать вам как начинающим аналитикам.

Вам выдаются результаты нескольких опросов.

1. Необходимо для каждой строки понять, является ли результат спамовым или настоящим. Результаты анализа необходимо прислать на почту вместе с решением.
2. Какими общими характеристиками обладают спамовые записи? Как часто они происходят?