Домашнее задание 3.3 Анализ данных

Выполнил студент Свиридов Иван СКБ182

Экспоненциальное распределение

[Ссылка на источник данных (https://www.kaggle.com/jessemostipak/hotel-booking-demand)](https://www.kaggle.com/jessemostipak/hotel-booking-demand)

Импортируем библиотеки:

In [48]:
```python
from functools import reduce
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math
```

Выведем наш dataset:

In [6]:

```
apparts = pd.read_csv('hotel_bookings.csv')
apparts
```

Out[6]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_nun
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	
2	Resort Hotel	0	7	2015	July	
3	Resort Hotel	0	13	2015	July	
4	Resort Hotel	0	14	2015	July	
...	...	...	...	...	...	
119385	City Hotel	0	23	2017	August	
119386	City Hotel	0	102	2017	August	
119387	City Hotel	0	34	2017	August	
119388	City Hotel	0	109	2017	August	
119389	City Hotel	0	205	2017	August	

119390 rows × 32 columns

Введение:

Предположим, что мы владельцы отеля. И нам необходимо провести анализ данных для успешной работы отеля. Возьмем количество дней, которое проходит между датой бронированием и датой приезда в отель посетителя. Эту разницу мы можем найти в нашем Dataset в колонке под именем **"lead_time"**.

Найдем основные характеристики по нашим данным:

In [7]:
```python
aparts['lead_time'].describe().to_frame()
```
Out[7]:

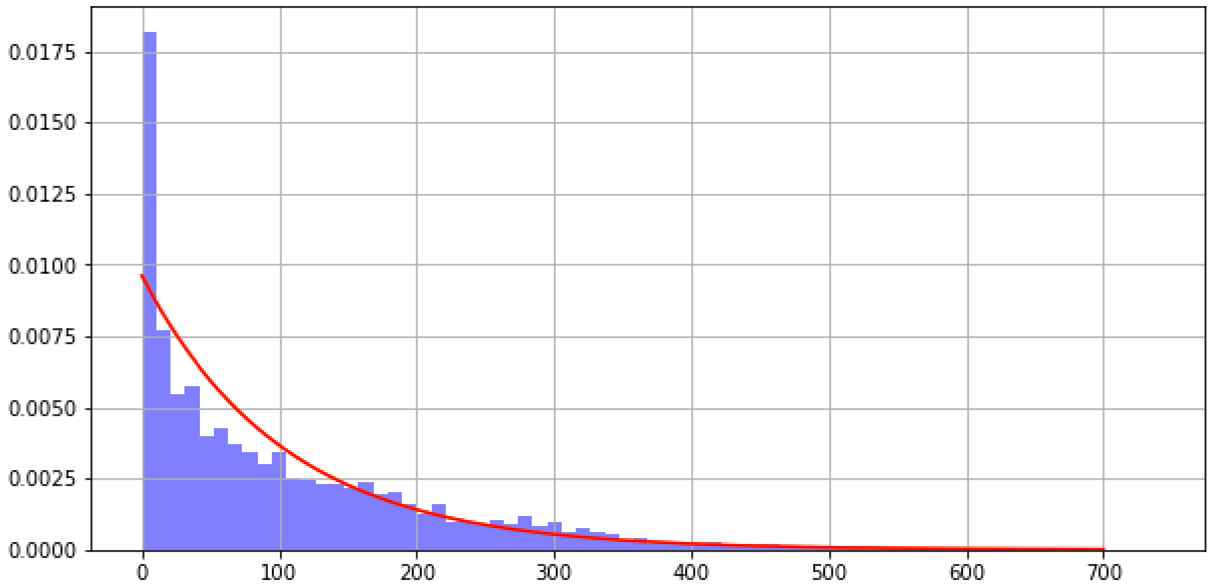
	lead_time
count	119390.000000
mean	104.011416
std	106.863097
min	0.000000
25%	18.000000
50%	69.000000
75%	160.000000
max	737.000000

Вариационный ряд:

In [14]:
```python
var_list=aparts['lead_time'].tolist()
var_list.sort()
print(var_list)
```
```
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
```

Построим график:

In [81]:
```
aparts['lead_time'].hist(bins=70, figsize=(10,5), alpha=0.5, facecolor='blue
x=np.linspace(0,700,100)
plt.plot(x, y:=(1/104)*np.exp((-1/104)*x))
plt.plot(x,y,'r')
plt.show()
```



На графике по оси X расположено количество дней, которое проходит между бронированием и приездом в отель. А по оси Y расположено количество случаев, которое соответсвует определенному количеству дней.

Как владельцы отеля, нам необходимо понимать через какое количество дней в среднем приезжают посетители в отель, после того как забронировали номер на сайте. Это очень важный показатель, чтобы получить наибольшую прибыль и отель был загружен максимально. Для нахождения этого параметра воспользуемся *выборочным средним*. И ниже опишем формулу и код на Python для нахождения этого параметра.

Выборочным моментом порядка k называется величина

$$\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Для краткости, выборочный момент первого прядка обозначают $\bar{X}$. $\hat{\alpha}_1$ называют *выборочным средним*, которое можно найти по формуле:

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

Вычислим выборочное среднее и медиану с помощью кода на языке Python:

- первый способ (с помощью вызова метода)

In [40]:
```
apparts['lead_time'].mean()
```

Out[40]: 104.01141636652986

- второй способ (с помощью пошаговых вычислений)

In [42]:
```
num_el=aparts.shape[0] #количество строк
sum_el=aparts['lead_time'].sum() #сумма элементов в столбце
sr_mean=sum_el/num_el
print(sr_mean)
```

104.01141636652986

Из полученных вычислений сделаем вывод. Что в среднем, проходит *104 дня* между датой бронирования и приездом или почти *3,5 месяца*. И в принципе, за это время могут жить другие посетители, а также можно подготовить штат сотрудников и необходимую инфраструктуру для этих заездов.

Найдем выборочную дисперсию. И опишем формулу и код на Python для нахождения:

Выборочным центральным моментов порядка k называется величина, определенная формулой:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\alpha}_1)^k$$

$\hat{\mu}_2$ называют *выборочной дисперсией*, которое можно найти по формуле:

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\alpha}_1)^2$$

In [46]:
```
vr_list=[]
summ=0

for item in var_list:
    summ=summ+(item-sr_mean)**2
dis=summ/num_el
print(dis)
```

11419.625860117929

Среднее квадратичное отклонение:

Средним квадратичным отклонением случайной величины *X* называется арифметическое значение корня квадратичного из ее дисперсии.

$$\sigma(x) = \sqrt{D(x)}$$

In [51]:
```
math.sqrt(dis)
```
Out[51]: 106.86264950916166

Оптимальная оценка параметра λ:

Для нашего экспоненциального распределения с параметром λ, за оцениваемый параметр возьмем $\tau(\theta) = \frac{1}{\lambda}$. Предположим оценку для $\tau(\theta)$, которую найдем при помощи метода моментов:

$M_\theta X_1 = \frac{1}{\lambda} \to \lambda = \frac{1}{M_\theta X_1}$, тогда $\theta = \frac{1}{\bar{X}} \to \tau(\theta) = \hat{\theta} = \bar{X}$

Если применить метод максимального правдоподобия, то мы получим такую же оценку, проверим и убедимся в этом:

$L(\theta, x) = \prod\limits_{j=1}^{n} \theta e^{-\theta x_j} = (\theta)^n e^{-\theta \sum_{j=1}^n x_j} = 0$, тогда запишем условие максимума: $\frac{dL(\hat{\theta},x)}{d\theta} = 0$, и в итоге получаем: $n(\theta)^{n-1} e^{-\theta \sum_{j=1}^n x_j} - (\theta)^n \sum_{j=1}^n x_j e^{-\theta \sum_{j=1}^n x_j} = 0$, откуда получаем $\theta = \frac{1}{\bar{X}} \to \tau(\theta) = \hat{\theta} = \bar{X}$

Наша оценка параметра λ совпала при помощи метода моментов и метода максимального правдоподобия.

Приведем значение преложенной оценки λ:

Как мы выяснили выше, наш параметр λ равен единице деленной на выборочную дисперсию ($\frac{1}{barX}$). Покажем это в числах. Возьмем выборочное среднее, которые мы нашли выше $\bar{X} = 104$, тогда $\lambda = \frac{1}{104}$

Найдем значение теоретического математического ожидания:

Воспользуемся формулой для нахождения математического ожидания для экспоненциального распределения.

$M\xi = \int_{-\infty}^{\infty} xf(x)dx = \lambda \int_0^{\infty} xe^{\lambda x} dx = -\lim\limits_{b\to\infty}((b + \frac{1}{\lambda})e^{-\lambda b})$

$\to (0 - (0 + \frac{1}{\lambda}e^0) = -(0 - \frac{1}{\lambda}) = \frac{1}{\lambda}$

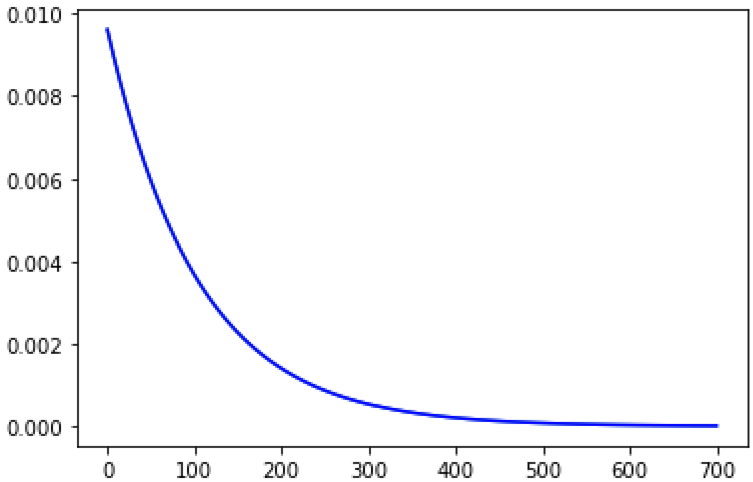
Подставим значение λ, тогда $M\xi = 104$

Найдем значение теоретической дисперсии:

Найдем теоретический график плотности:

In [78]:

```
x=np.linspace(0,700,100)
plt.plot(x, y:=(1/104)*np.exp((-1/104)*x))
plt.plot(x,y,'b')
plt.show()
```



In [ ]:

```

```