

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ)"**

ЖУРНАЛ ПРАКТИКИ

Студента 2 курса

Гординского Дмитрия Михайловича

Институт №8 «Информационные технологии и прикладная математика»

Кафедра №804 «Теория вероятностей и компьютерное моделирование»

Учебная группа М8О-204Б-20

Направление 01.03.04

Прикладная математика

Вид практики Учебная (вычислительная) в Московском Авиационном Институте(НИУ)

Руководитель практики от МАИ Зайцева О.Б.

Гординский Д.М /

/ 11 июля 2022 г.

Москва, 2022

1. Место и сроки проведения практики

Дата начала практики 29 июня 2022 г.

Дата окончания практики 11 июня 2022 г.

Наименование предприятия МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ(НИУ)

Название структурного подразделения Кафедра 804

2. Инструктаж по технике безопасности

_____/_____/ 29 июня 2022 г.

3. Индивидуальное задание студенту

1. Разобраться с теорией.
2. Привести пример решения задачи.
3. Написать отчет.

4. План выполнения индивидуального задания

1. Изучить теорию по Моделям выживаемости.
2. Ознакомиться с необходимыми библиотеками для работы с данными и их графическим представлением.
3. Решить задачу по анализу данных с применением методов анализа выживаемости.

Руководитель практики от МАИ: _____/_____/

_____/_____/ 29 июня 2022 г.

5. Отзыв руководителя практики

Задание на практику выполнено в полном объеме. Материалы, изложенные в отчете студента, полностью соответствуют индивидуальному заданию. Рекомендую оценку отлично.

Руководитель _____/_____/ 11 июля 2022 г.

Отчет студента

Содержание

1	Что такое “Анализ выживаемости”?	3
1.1	Основные понятия	3
1.1.1	Функция выживания (Survival function)	3
1.1.2	Функция риска (Hazard function)	3
1.1.3	Цензурирование (censoring)	4
1.1.4	Медиана ожидаемого времени жизни (median number of survival days)	4
1.1.5	Доверительный интервал (confidence interval)	4
1.1.6	Усечение (truncation)	4
1.2	Непараметрические методы оценивания распределения длительностей	4
1.2.1	Оценка Каплана — Мейера	5
1.2.2	Оценка Нельсона — Аалена	5
2	Пример решения задачи	6
2.1	Проанализируем данные пола:	6
2.2	Применяем <i>Оценку Каплана — Мейера</i>	7
2.2.1	Сделаем таблицу событий	7
2.2.2	Найдем вероятность выживания для каждого момента времени и вероятность с доверительным интервалом	8
2.2.3	Найдем медиану времени выживания	9
2.2.4	Найдем вероятность смерти для $\forall t$	10
2.3	Применяем <i>Оценку Нельсона — Аалена</i>	11
2.3.1	Найдем риск для каждого момента времени и риск с доверительным интервалом	11
2.3.2	Сравним кумулятивную функцию риска и кумулятивную плотность (<i>вероятность смерти</i>):	12
2.4	Анализ выживания для групп	12
2.4.1	Найдем вероятность выживания среди мужчин и женщин	13
2.4.2	Сравним кумулятивную плотность выживания с кумулятивной функцией риска	15
2.5	Итоги	16
3	Вывод	16

1 Что такое “Анализ выживаемости”?

Анализ выживаемости — набор статистических моделей, благодаря которым можно оценить вероятность наступления того или иного события. Анализ занимается моделированием процессов наступления *интересующих* нас (критических) событий для элементов той или иной совокупности (изначально — «смерти» для элементов совокупности живых существ).

Интересным событием может быть что угодно. Это может быть фактическая смерть, рождение, выход на пенсию и т. д.

Название “*survival analysis*” взято из медицины, т.к. цель анализа заключается в изучении продолжительности жизни пациента после приема препарата или других факторов влияния на здоровье.

1.1 Основные понятия

1.1.1 Функция выживания (Survival function)

Пусть T — неотрицательная случайная величина, представляющая собой время ожидания до наступления некоторого события. Для простоты будем использовать терминологию анализа выживаемости, называя исследуемое событие «смертью», а время ожидания — временем «выживания»

Функция выживания сопоставляет некоторому числу t вероятность того, что случайная величина T примет значение, не меньшее t . Иначе говоря, это вероятность того, что некоторое состояние «проживет» как минимум t единиц времени:

$$S(t) = \mathbb{P}\{T > t\} = 1 - \mathbb{P}\{T \leq t\}$$

Например, если мы хотим знать, какова вероятность того, что безработный индивид не сможет найти работу в течение полугода после начала поиска, то достаточно рассмотреть функцию выживания для $t = 6$ месяцев.

1.1.2 Функция риска (Hazard function)

Функцию риска можно охарактеризовать как вероятность того, что событие произойдет за бесконечно малый интервал времени при условии, что оно не произошло к моменту времени t .

$$h(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + dt | T \geq t)}{dt}$$

Числитель этого выражения — условная вероятность того, что событие произойдет в интервале $(t, t + dt)$, если оно не произошло ранее, а знаменатель — ширина интервала. Разделив одно на другое, получаем интенсивность осуществления события в единицу времени. Устремляя ширину интервала к нулю и переходя к пределу, получаем *мгновенную интенсивность осуществления события*.

Т. к. вышесвязанные функции связаны друг с другом, можно показать, что:

$$S(t) = \exp\left(-\int_0^t h(x)dx\right)$$

Интеграл в фигурных скобках в этом уравнении называют *кумулятивным риском* и обозначают как:

$$H(t) = \int_0^t h(x)dx$$

Можно рассматривать $H(t)$ как сумму всех рисков при переходе от момента времени 0 к t .

1.1.3 Цензурирование (censoring)

Цензурирование — вид неполноты информации, при котором наблюдения не содержат точной длительности изучаемого состояния. Различают цензурирование справа, слева и интервальное:

1. Цензурировано справа — о наблюдаемом состоянии известно лишь, что оно продлилось не менее определенного времени.
2. Цензурировано слева — о состоянии известно лишь, что оно продлилось не более определенного времени.
3. На интервале — известны только границы длительности.

1.1.4 Медиана ожидаемого времени жизни (median number of survival days)

Это точка на временной оси, в которой кумулятивная функция выживания равна 0,5.

Другими словами, *медиана* — время, выраженное в месяцах или годах, когда ожидается, что половина пациентов будет жива. Это означает, что шанс выжить после этого времени составляет 50 процентов.

1.1.5 Доверительный интервал (confidence interval)

Доверительный интервал — интервал, который покрывает неизвестный параметр с заданной надёжностью. Вероятность, с которой в условиях данного эксперимента полученные экспериментальные данные можно считать надёжными (достоверными), называют доверительной вероятностью или надёжностью. Величина доверительной вероятности определяется характером производимых измерений. Мы будем считать доверительную вероятность равной 95 %.

1.1.6 Усечение (truncation)

Усечением, или урезанием, называется вид неполноты информации, при котором какая-то область возможных значений длительности оказывается недостаточно представленной в выборке: состояния, длительность которых слишком велика или, наоборот, слишком мала, просто не включаются в анализируемые данные. В нашей задаче мы будем называть их (removed) — пациенты, которые больше не являются частью нашего эксперимента. Если человек умирает или подвергается цензуре, то он попадает в эту категорию.

1.2 Непараметрические методы оценивания распределения длительностей

При отсутствии цензурирования и усечения для оценивания закона распределения вероятностей может использоваться эмпирическая функция распределения, из которой легко получить оценки для других характеристик случайной величины: survival function etc. Но в нашем случае это невозможно, т. к. мы имеем дело с неполнотой данных. Эту проблему решают непараметрические методы оценки.

1.2.1 Оценка Каплана — Мейера

Оценка Каплана-Мейера — это непараметрическая статистика, используемая для оценки функции выживания на основе данных о жизни. В медицинских исследованиях он часто используется для измерения доли пациентов, живущих в течение определенного времени после лечения или постановки диагноза. Например: подсчет количества времени, которое прожил конкретный пациент после того, как у него был диагностирован рак или началось его лечение.

$$\hat{S}(t) = \prod_{t_j \leq t} \frac{n_j - d_j}{n_j}$$

где

$\hat{S}(t)$ = Вероятность того, что испытуемый жив в момент времени t

n_j = Количество испытуемых, оставшихся в живых непосредственно перед моментом времени t_j

d_j = Количество событий в момент времени t_j

Можем переписать формулу выше так:

$$S(t_j) = S(t_{j-1}) \left(1 - \frac{d_j}{n_j}\right)$$

где

$S(t_j)$ = Вероятность того, что испытуемый жив в момент времени t_j

n_j = Количество испытуемых, оставшихся в живых непосредственно перед моментом времени t_j

d_j = Количество событий в момент времени t_j

$S(0) = 1$

$t_0 = 0$

1.2.2 Оценка Нельсона — Аалена

Мы можем визуализировать совокупную информацию о выживании, используя функцию риска *Нельсона-Аалена* $h(t)$. Функция риска $h(t)$ дает нам вероятность того, что субъект, находящийся под наблюдением в момент времени t , имеет интересующее событие (смерть) в это время. Чтобы получить информацию о функции риска, мы не можем преобразовать оценку Каплана-Мейера. Для этого существует соответствующая непараметрическая оценка кумулятивной функции риска:

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j}$$

где

$\hat{H}(t)$ = Кумулятивная вероятность риска

n_j = Количество испытуемых, оставшихся в живых непосредственно перед моментом времени t_j

d_j = Количество событий в момент времени t_j

2 Пример решения задачи

В качестве примера для анализа выживаемости возьмем заболевание *Chronic Granulomatous Disease*

Хроническая гранулематозная болезнь (ХГБ) [CGD] — это наследственное заболевание, которое возникает, когда тип лейкоцитов (фагоцитов), которые обычно помогают организму бороться с инфекциями, не работает должным образом. В результате фагоциты не могут защитить организм от бактериальных и грибковых инфекций. У людей с хронической гранулематозной болезнью могут развиваться инфекции в легких, коже, лимфатических узлах, печени, желудке и кишечнике или других областях. У них также могут образовываться скопления лейкоцитов в зараженных областях. У большинства людей ХГБ диагностируется в детстве, но у некоторых людей диагноз может не ставиться до зрелого возраста.

В датасете нас интересует:

- время от начала наблюдения за пациентом до события (смерти) ($tstop - tstart$)
- пол пациента (sex)
- $status = \{status == 0 = \text{alive}, status == 1 = \text{dead}\}$

В дальнейшем может пригодиться возраст (age) для анализа выживаемости по группам.

2.1 Проанализируем данные пола:

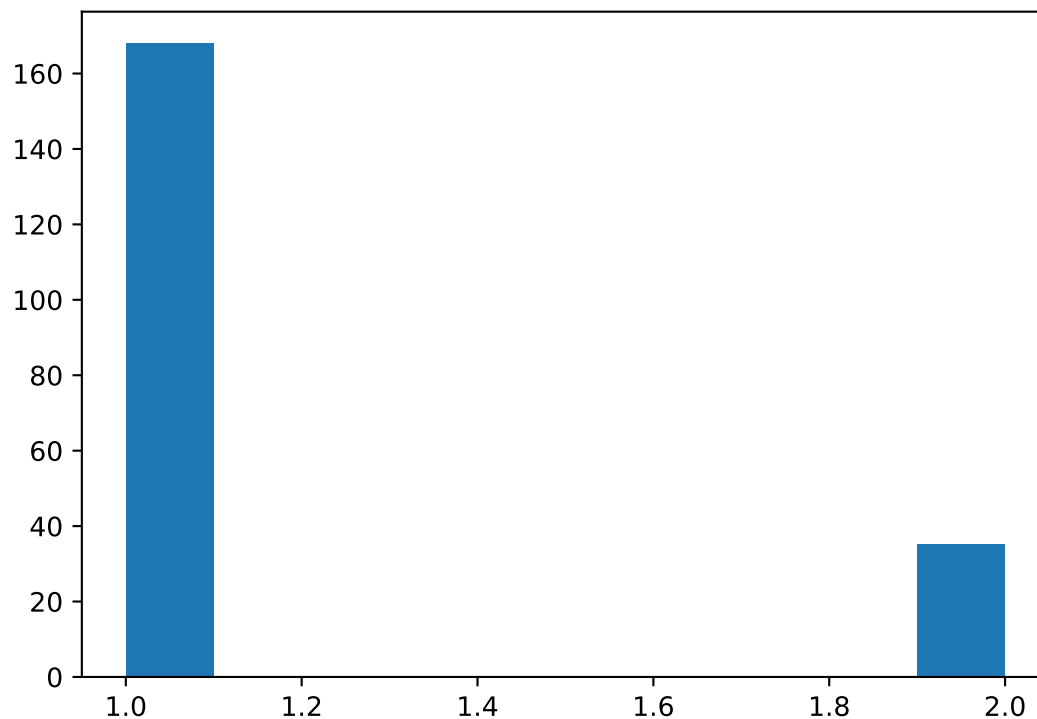
для начала подключим необходимые библиотеки...

lifelines — содержит необходимые нам методы для исследования вероятностей и времени жизни.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from lifelines import KaplanMeierFitter
from lifelines import NelsonAalenFitter
# считываем данные
data = pd.read_csv("cgd.csv")
head = data.head()
```

Unnamed: 0	id	center	random	treat	sex	age	...	steroids	propylac	hos.cat	tstart	enum	tstop	status	
0	1	1	Scripps Institute	1989-06-07	rIFN-g	2	12	...	0	0	US:other	0	1	219	1
1	2	1	Scripps Institute	1989-06-07	rIFN-g	2	12	...	0	0	US:other	219	2	373	1
2	3	1	Scripps Institute	1989-06-07	rIFN-g	2	12	...	0	0	US:other	373	3	414	0
3	4	2	Scripps Institute	1989-06-07	placebo	1	15	...	0	1	US:other	0	1	8	1
4	5	2	Scripps Institute	1989-06-07	placebo	1	15	...	0	1	US:other	8	2	26	1

```
data.loc[data.sex == "male", "sex"] = 1
data.loc[data.sex == "female", "sex"] = 2
plt.hist(data["sex"]) #гистограмма полов
plt.show()
```



2.2 Применяем Оценку Каплана — Мейера

```
kmf = KaplanMeierFitter()
# в нашем случае "status" === "dead"
data.loc[:, "time"] = data.loc[:, "tstop"] - data.loc[:, "tstart"] # time=tstop-tstart
kmf.fit(durations = data["time"], event_observed = data["status"])
```

```
## <lifelines.KaplanMeierFitter:"KM_estimate", fitted with 203 total observations, 127 right-
censored observations>
```

2.2.1 Сделаем таблицу событий

Нам это нужно, чтобы отделить цензурированные данные, получить необходимые временные данные для применения методов оценки.

```
print(kmf.event_table)
```

```
##          removed  observed  censored  entrance  at_risk
## event_at
## 0.0             0         0         0         203      203
## 2.0             1         1         0          0      203
## 4.0             3         2         1          0      202
## 5.0             1         1         0          0      199
## 6.0             1         1         0          0      198
```



```
## ...      ...      ...      ...      ...
## 371.0      1      0      1      0      7
## 373.0      2      1      1      0      6
## 376.0      1      0      1      0      4
## 382.0      1      0      1      0      3
## 388.0      2      0      2      0      2
##
## [154 rows x 5 columns]
```

где

- *event_at* — хранит значение временной шкалы для нашего набора данных. т. е. когда пациент наблюдался в нашем эксперименте или когда был проведен эксперимент, хранит значение дней выживания для субъектов.
- *at_risk* — хранит количество текущих пациентов, находящихся под наблюдением.

$$at_risk = currentpatientsat_risk + entrance - removed$$

- *entrance* — хранит значение новопришедших пациентов. Т. е. во время проведения эксперимента появлялись новые больные.
- *censored* — если человек все еще жив по окончании эксперимента, то мы добавляем его в эту категорию.
- *observed* — содержит количество умерших пациентов во время эксперимента.
- *removed* — содержит количество пациентов, которые “выпадают” из эксперимента
 $removed = observed + censored$

2.2.2 Найдем вероятность выживания для каждого момента времени и вероятность с доверительным интервалом

Для начала найдем вероятность выживания за время t
(см. раздел 1.2.1)

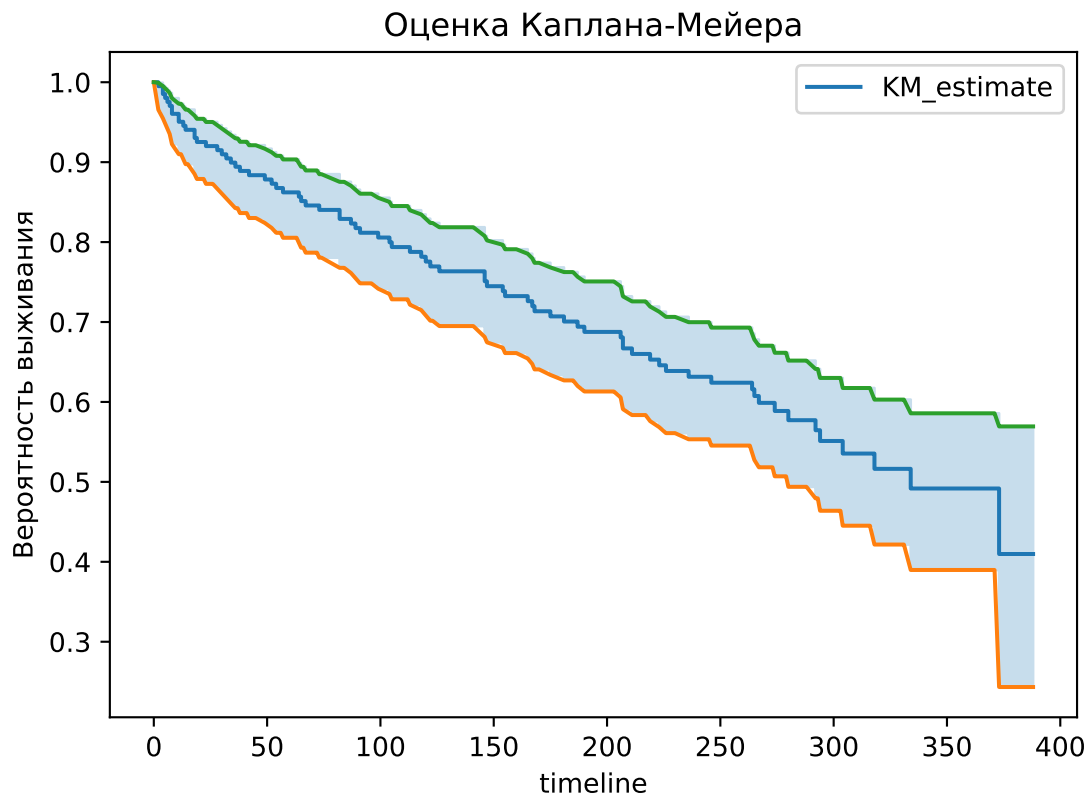
Не ограничивая общности, $\forall t$: Пусть $t = 6 \Rightarrow$

```
# Вероятность выживания после 6 дней
e0 = kmf.event_table.iloc[0, :]
e2 = kmf.event_table.iloc[1, :]
e4 = kmf.event_table.iloc[2, :]
e6 = kmf.event_table.iloc[3, :]
s0 = (e0.at_risk - e0.observed)/e0.at_risk
s2 = (e2.at_risk - e2.observed)/e2.at_risk
s4 = (e4.at_risk - e4.observed)/e4.at_risk
s6 = (e6.at_risk - e6.observed)/e6.at_risk
s6 = s0 * s2 * s4 * s6
print(s6)
```

```
## 0.9802708121890239
```

Найдем вероятность выживания $\forall t$:

```
kmf.survival_function_  
plt.title("Оценка Каплана-Мейера")  
plt.ylabel("Вероятность выживания")  
kmf.plot()  
csf = kmf.confidence_interval_survival_function_  
plt.plot(csf["KM_estimate_lower_0.95"], label="lower")  
plt.plot(csf["KM_estimate_upper_0.95"], label="upper")  
plt.show()
```



По графику видно, что с течением времени вероятность выживания уменьшается.

2.2.3 Найдем медиану времени выживания

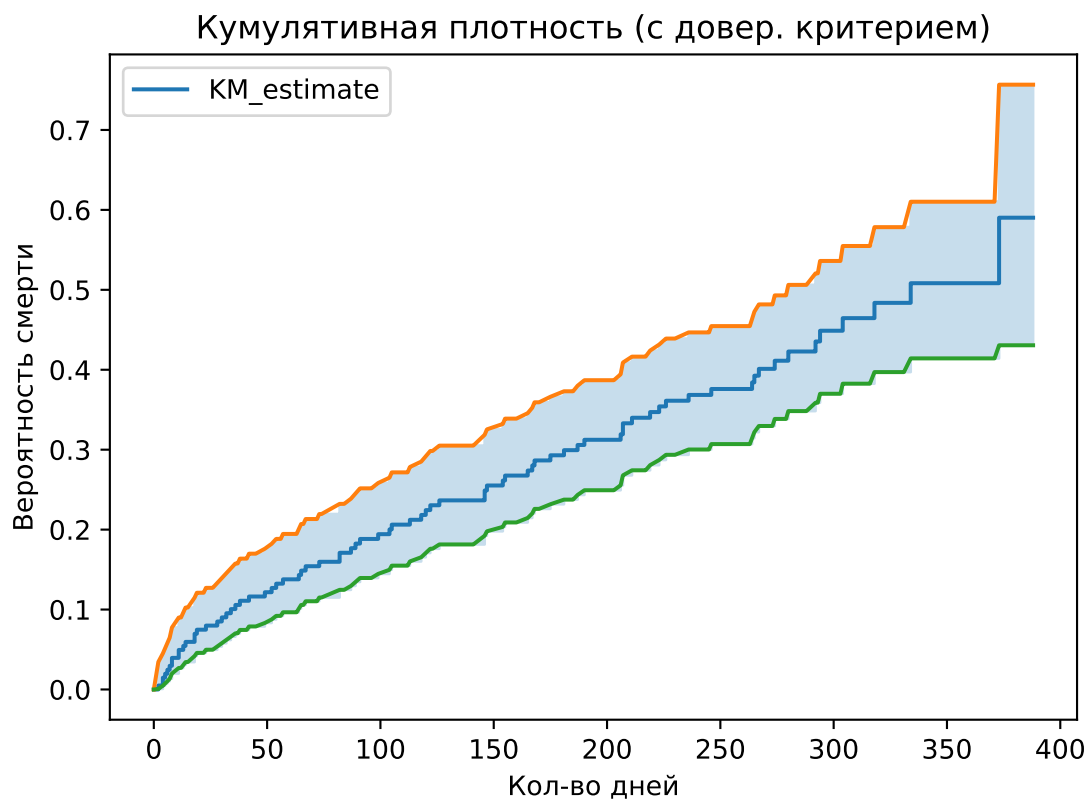
```
print("Медиана времени выживания", kmf.median_survival_time_)
```

```
## Медиана времени выживания 334.0
```

2.2.4 Найдем вероятность смерти для $\forall t$

Сделаем график кумулятивной функции плотности и кумулятивной плотности с доверительным критерием

```
kmf.plot_cumulative_density()
ccf = kmf.confidence_interval_cumulative_density_
plt.plot(ccf["KM_estimate_lower_0.95"], label="lower")
plt.plot(ccf["KM_estimate_upper_0.95"], label="upper")
plt.title("Кумулятивная плотность (с довер. критерием)")
plt.xlabel("Кол-во дней")
plt.ylabel("Вероятность смерти")
plt.show()
```



Видим полностью обратный график к вероятности выживания, что неудивительно, ведь кумул. ф-я плотности:

$$F(t) = 1 - S(t)$$

2.3 Применяем *Оценку Нельсона — Аалена*

```
naf = NelsonAalenFitter()
naf.fit(durations = data["time"], event_observed = data["status"])
```

```
## <lifelines.NelsonAalenFitter:"NA_estimate", fitted with 203 total observations, 127 right-
censored observations>
```

2.3.1 Найдем риск для каждого момента времени и риск с доверительным интервалом

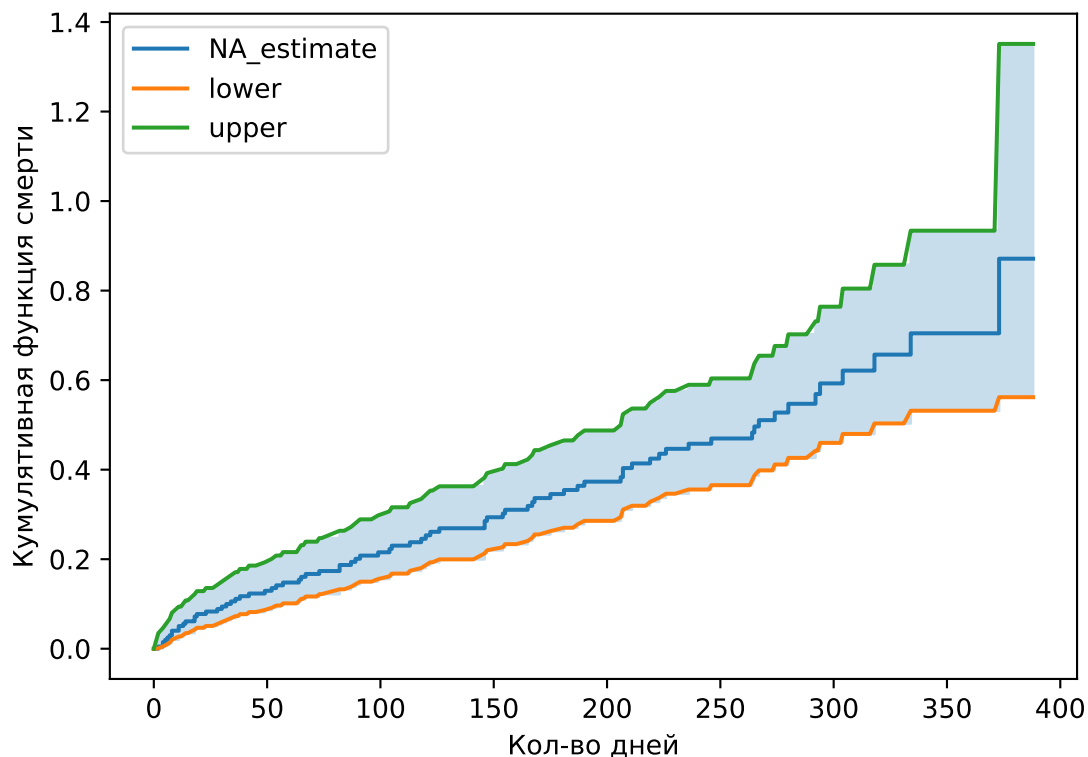
По нашей таблице событий (см. раздел 2.2.1) мы считаем функцию риска (см. раздел 1.2.2)

Также можем предсказать значение функции риска для $\forall t$

```
print("300 дней: ", naf.predict(300))
```

```
## 300 дней: 0.5927191310947535
```

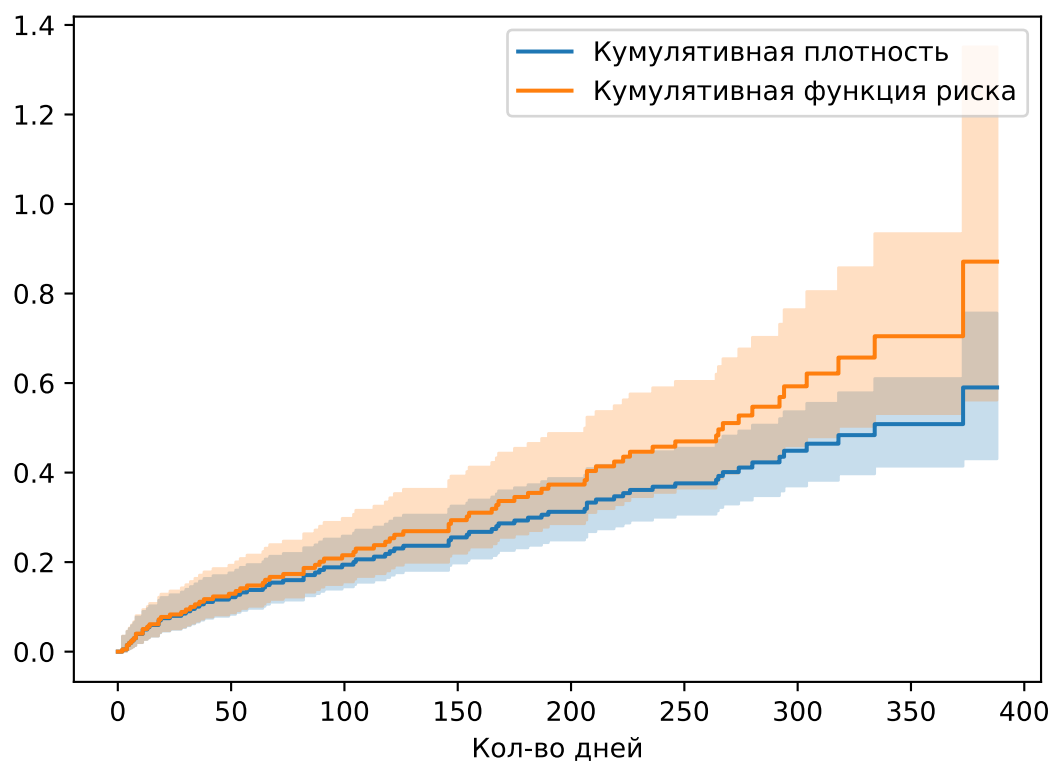
```
naf.plot_cumulative_hazard()
ci = naf.confidence_interval_
plt.plot(ci["NA_estimate_lower_0.95"], label="lower")
plt.plot(ci["NA_estimate_upper_0.95"], label="upper")
plt.xlabel("Кол-во дней")
plt.ylabel("Кумулятивная функция смерти")
plt.legend()
```



Другими словами, функция риска измеряет *общую сумму риска*, накопленного к моменту времени t

2.3.2 Сравним кумулятивную функцию риска и кумулятивную плотность (вероятность смерти):

```
kmf.plot_cumulative_density(label="Кумулятивная плотность")
naf.plot_cumulative_hazard(label="Кумулятивная функция риска")
plt.xlabel("Кол-во дней")
plt.show()
```



2.4 Анализ выживания для групп

Сначала сравним выживаемость мужчин и женщин:

```
kmfm = KaplanMeierFitter() # мужчины
kmff = KaplanMeierFitter() # женщины

data.loc[data.sex == "male", "sex"] = 1
data.loc[data.sex == "female", "sex"] = 2

male = data.query("sex == 1")
female = data.query("sex == 2")
```

```

kmfm.fit(durations=male["time"], event_observed=male["status"], label="male")
kmff.fit(durations=female["time"], event_observed=female["status"], label="female")

# сделаем таблицы событий отдельно для м. и ж.
kmfm.event_table
kmff.event_table
print(kmfm.event_table.head())
print(kmff.event_table.head())

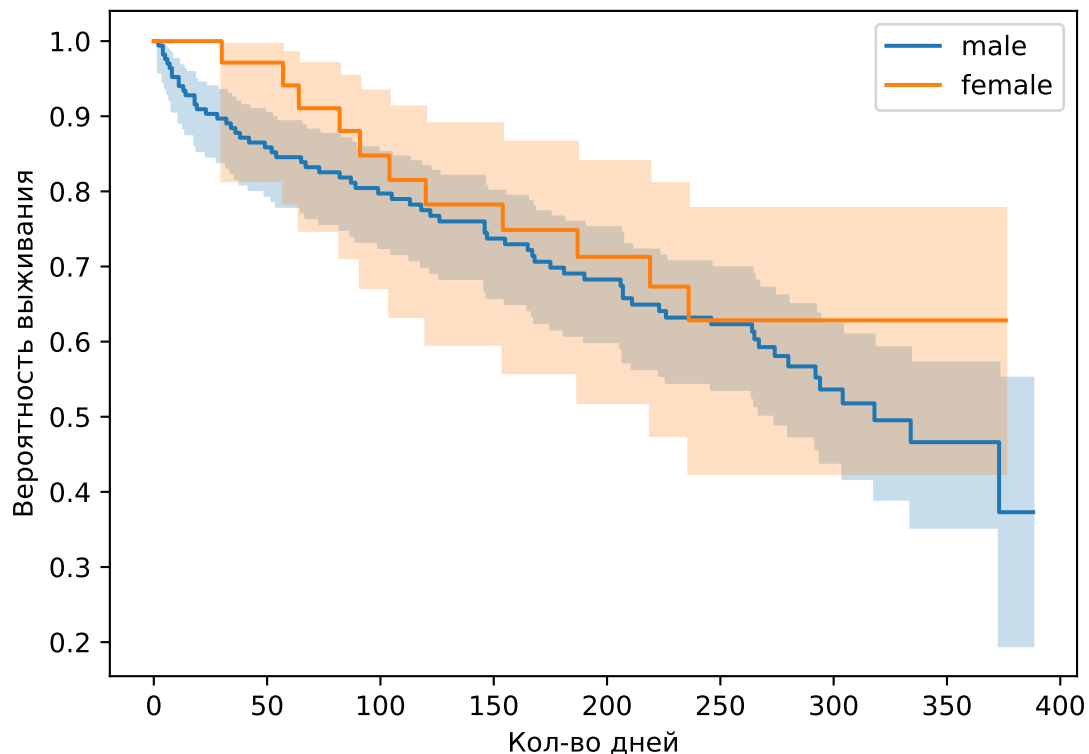
```

2.4.1 Найдем вероятность выживания среди мужчин и женщин

```

kmfm.survival_function_
kmff.survival_function_
kmfm.plot()
kmff.plot()
plt.xlabel("Кол-во дней")
plt.ylabel("Вероятность выживания")
plt.show()

```



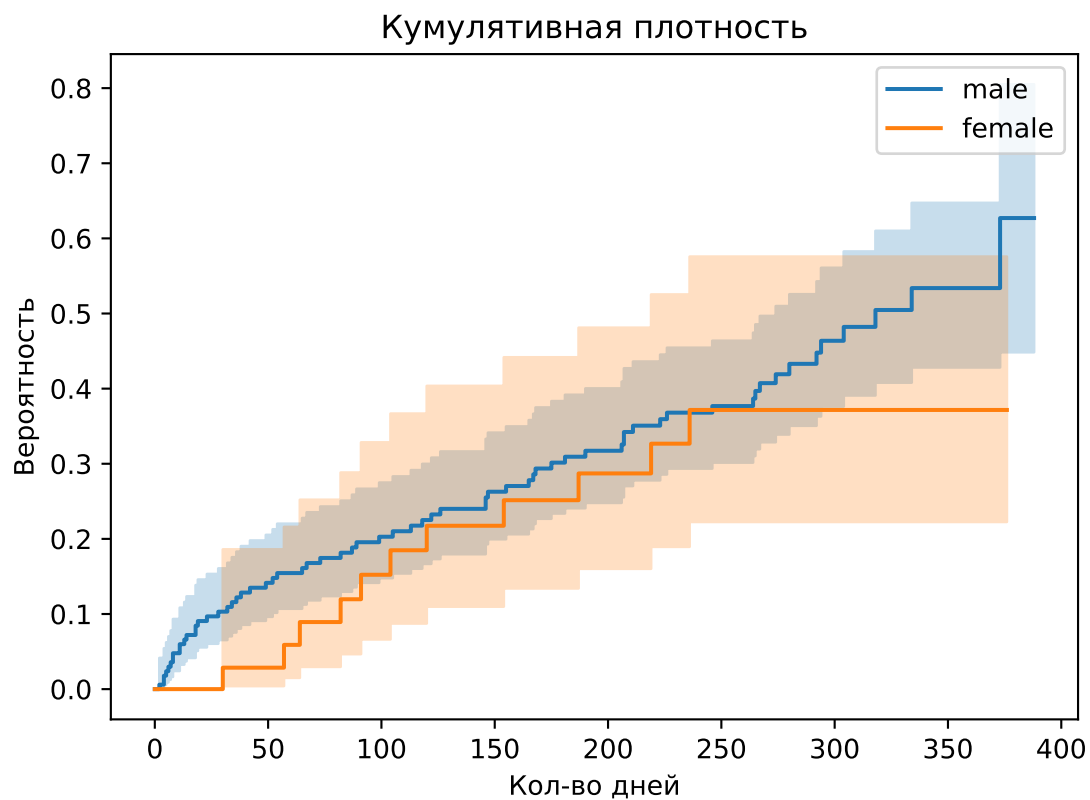
Несложно заметить, что вероятность выживания женщин выше, чем у мужчин. Однако такие выводы не совсем точны, т.к. мужчин много больше, чем женщин. Но до дня ~250 все равно вероятность у женщин выше.

Получается, что кумулятивная плотность будет ниже у женщин

```

kmfm.plot_cumulative_density()
kmff.plot_cumulative_density()
plt.title("Кумулятивная плотность")
plt.xlabel("Кол-во дней")
plt.ylabel("Вероятность")
plt.show()

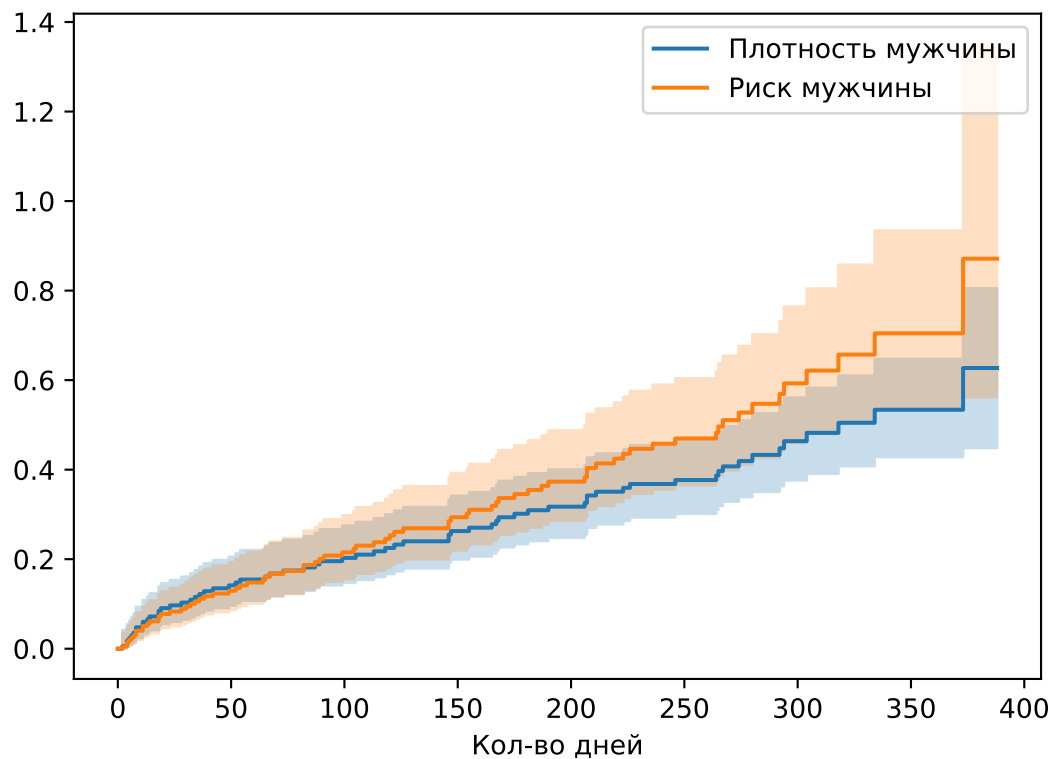
```



2.4.2 Сравним кумулятивную плотность выживания с кумулятивной функцией риска

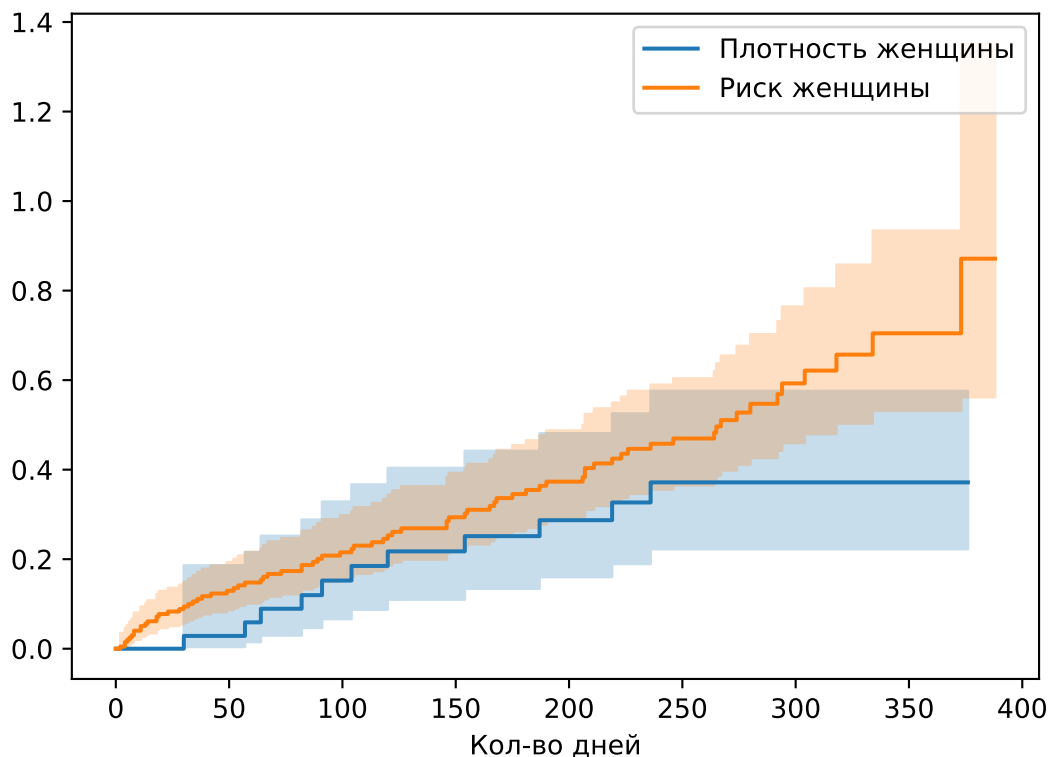
мужчины:

```
nafm = NelsonAalenFitter()
nafm.fit(durations = data["time"], event_observed = data["status"])
kmfm.plot_cumulative_density(label="Плотность мужчины")
nafm.plot_cumulative_hazard(label="Риск мужчины")
plt.xlabel("Кол-во дней")
plt.show()
```



женщины:

```
naff = NelsonAalenFitter()
naff.fit(durations = data["time"], event_observed = data["status"])
kmff.plot_cumulative_density(label="Плотность женщины")
naff.plot_cumulative_hazard(label="Риск женщины")
plt.xlabel("Кол-во дней")
plt.show()
```

⇒с течением времени риск увеличивается.

2.5 Итоги

Проанализировав выживаемость после заболевания CGD, можно сказать, что примерно за 1 год вероятность выживания становится равной 0. ⇒чем больше времени проходит, тем больше смертельный риск.

Анализ показал, что на нашей выборке у женщин немного больше шансов на выживание, нежели у мужчин. Анализировать действие лечебных препаратов, которые давали пациентам, не стал, т.к. это исключительно профилактическое-консервативное лечение, на пациентов с данным диагнозом практически не влияет.

3 Вывод