

Генеральная совокупность и выборка. Выборочные оценки. Статистики.

- Генеральная совокупность и выборки из нее;
- Числовые характеристики распределения СВ: математическое ожидание
- Начальные и центральные моменты как характеристики СВ;
- Статистики как функции векторной СВ, выборочные оценки;
- Свойства статистик: состоятельность, несмешенность, эффективность;

монета | МАТБАЗА

Генеральная совокупность и выборка.
Выборочные оценки. Статистики. Свойства статистик.

8 ноября
занятие 2.4.

Игорь Нехаев
Специалист по анализу данных
и машинному обучению

▼ Генеральная совокупность (ГС) и выборка из нее

Предположим, мы захотели узнать как оценивают по 10-балльной шкале работу компании наши клиенты среднего возраста, построить распределение СВ "оценка компании клиентом среднего возраста". Как быть? Если бы смогли провести опрос всех клиентов от 30 до 50 лет, то получили бы **генеральную совокупность** значений СВ и смогли бы на основе этой генеральной совокупности получить какое-то истинное распределение этой СВ.

Но что, если таких клиентов у нас более 1000? А если более 10000? Каждого опросить просто нереально. Не факт, что даже дозвонимся или достучимся. Ну и можем не получить ответ, даже если дозвонимся или достучимся по email.

Опр. Генеральной совокупностью для данной СВ назовем все мыслимое (конечное или бесконечное) множество значений СВ, которое мы могли бы получить при наблюдении данной СВ.

Имея такое множество значений, мы могли бы построить истинный закон распределения СВ. Т.е. можно считать, что:

| ЗНАНИЕ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ == ЗНАНИЕ ЗАКОНА
| РАСПРЕДЕЛЕНИЯ СВ

ВЫБОРКА

А можно ли опросить из всех только 50-100 клиентов и на основе этих данных построить распределение и его использовать вместо истинного? Насколько сильно данное эмпирическое распределение будет отличаться от истинного? (на прошедшем мастер-классе мы убедились, что несильно будет отличаться, но может это нам повезло?). Такая часть из генеральной совокупности (значения СВ, полученные для части объектов) называется **ВЫБОРКОЙ** или выборочными значениями.

И иногда даже нас интересует не все распределение, а только его некоторые характеристики - среднее и разброс значений СВ. И оценки таких характеристик на основе только выборки, как части ГС называются выборочными оценками. И вот они, к нашему счастью, подчиняются определенным распределениям (нормальному, хиквадрат). А значит, мы можем оценить их поведение и понять как обеспечить их устойчивость и приближение к истинным значениям, которые мы могли бы получить, зная всю генеральную совокупность.



ВОПРОС

Что можно считать ГС для СВ "Рост сотрудника компании"? Выберите наиболее точный с вашей точки зрения ответ:

1. Множество измерений ростов всех сотрудников компании на данный момент
2. Множество измерений ростов всех сотрудников компании, которое удалось измерить в данный день в офисе компании

3. Множество измерений ростов всех сотрудников компании, которые когда-либо работали в компании.
4. Множество измерений ростов всех сотрудников компании, которые когда-либо работали в компании, объединенное с множеством мыслимых измерений работников, которые в будущем будут работать на компанию.

▼ Математическое ожидание или генеральное среднее для функции СВ $g(X)$

Рассмотрим произвольную (непрерывную) функцию от СВ X : $g(X)$. Явно то, что она тоже является некоторой СВ $Y = g(X)$. Поэтому тоже подчиняется какому-то закону распределения и имеет какое-то среднее значение.

Пусть X - является ДСВ и принимает значения x_i с вероятностью $p_i = p(X = x_i)$, $i = 1..n$.

Опр. Математическим ожиданием функции $g(X)$ (или СВ Y) называется средневзвешенное значение $g(x)$, рассчитываемое по формуле:

$$(*) M(g(X)) = \sum_{i=1}^n p_i \cdot g(x_i)$$

Иногда говорят о мат.ожидании как о генеральном среднем СВ $g(X)$. Действительно, если бы нам были известны все значения из ГС, то формула была бы очень проста:

$$(**) M(g(X)) = \frac{1}{|\Gamma C|} \sum_{\forall x_i \in \Gamma C} g(x_i)$$

Действительно, мы могли бы просто вычислить средний рост сотрудников нашей компании по этой формуле, если бы нам был известен рост всех сотрудников компании. Этой формулой мы сможем воспользоваться только в случае, когда ГС конечна. В общем случае, нам надо будет идентифицировать закон распределения ($p_i = p(X = x_i)$ для ДСВ или $f_X(x)$ для НСВ) и использовать формулу (*).

Для НСВ X математическое ожидание СВ $g(X)$ находят с помощью интеграла и плотности распределения $f_X(x)$:

$$M(g(X)) = \int_{-\infty}^{+\infty} f_X(x) \cdot g(x) dx$$

▼ ПРИМЕР. Анализ времени подключения заявок

В качестве примера рассмотрим СВ X ="Время подключения клиента по заявке, поданной в 2023 году".

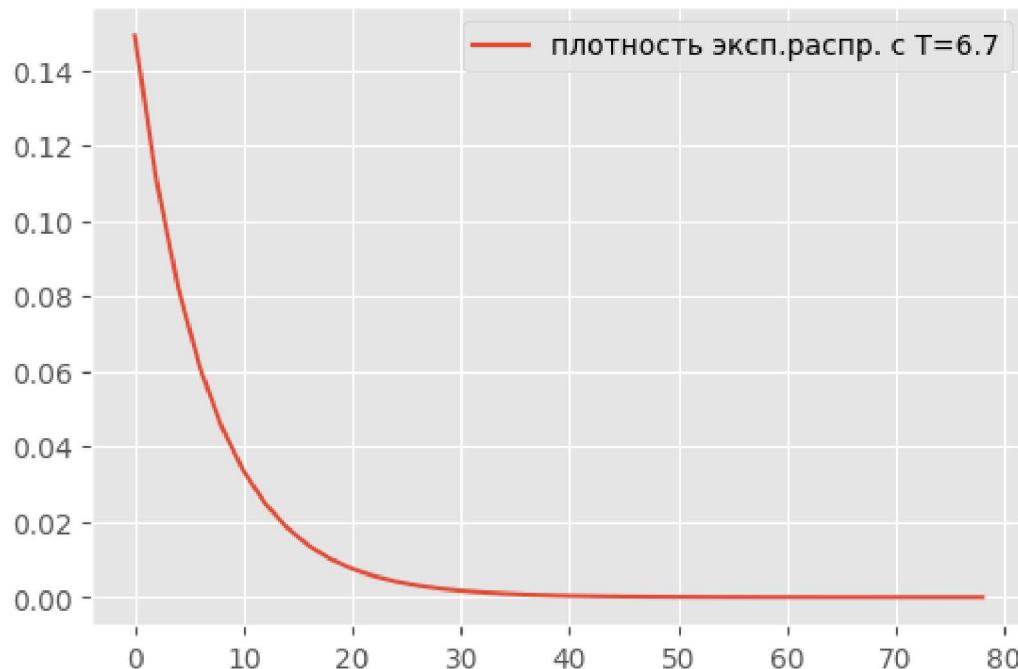
Мы знаем информацию по всем заявкам, которые зарегистрированы у нас. Это генеральная совокупность или нет?

Восстановим получившуюся плотность распределения времени (см предыдущий урок)

```
import numpy as np
np.set_printoptions(precision=3)
import scipy.stats as stat
import matplotlib.pyplot as plt
plt.style.use('ggplot')

Tmean = 6.7
exp_val = stat.expon(scale=Tmean)

# нарисуем плотность распределения интервала времени
xbins = np.arange(0, 80, 2)
plt.figure(figsize=(6, 4))
plt.plot(xbins, exp_val.pdf(xbins), label=f"плотность эксп.распр. с T={Tmean}")
plt.legend()
plt.show()
```



```
# найдем мат.ожидание от X и от X**2
from scipy import integrate

# зададим подинтегральные функции
f_X = lambda x: exp_val.pdf(x) * x
f_X2 = lambda x: exp_val.pdf(x) * x**2

# найдем M(X) - выполним интегрирование с помощью пакета
result = integrate.quad(f_X, -np.inf, +np.inf)
M_X = result[0]

# посмотрим результат
print(result)

(6.70000000000001, 1.3342089496107455e-09)
```

собственно, мы этого и должны были ожидать, ведь мы использовали именно параметр "среднее время подключения" для задания плотности распределения. Найдем теперь мат.ожидание других интересных функций от X.

```
# найдем M(X**2) - выполним интегрирование с помощью пакета
result = integrate.quad(f_X2, -np.inf, +np.inf)
M_X2 = result[0]

# посмотрим результат
print(M_X2)
```

89.78

```
print(f"Мат.ожидание X M(X) = {round(M_X, 3)}")
print(f"Квадрат от мат.ожидания X M(X)**2 = {round(M_X**2, 3)}")
print(f"Мат.ожидание X**2 M(X**2) = {M_X2}")
```

Мат.ожидание X M(X) = 6.7
 Квадрат от мат.ожидания X M(X)**2 = 44.89
 Мат.ожидание X**2 M(X**2) = 89.78

Это интересно, бесконечный хвост плотности распределения "утащил" среднее (мат.ожидание) квадрата X далеко от квадрата среднего X.

А теперь добавим перчика в задачку. Предположим, что за каждого подключенного по заявке клиента компания выписывает премию, которую руководитель отдела Подключения раскидывает по участникам подключения :). В зависимости от скорости подключения премия составляет

$$bonus(x) = 500 + \frac{1000}{x} \text{ (руб)}$$

Т.е. подключили за 1 день - получите +1500 руб. Подключили за 2 дня - получите 1000 руб., ...

Оценим математическое ожидание премии за 1 заявку (в среднем сколько за одну заявку получаем при таком распределении времени подключения):

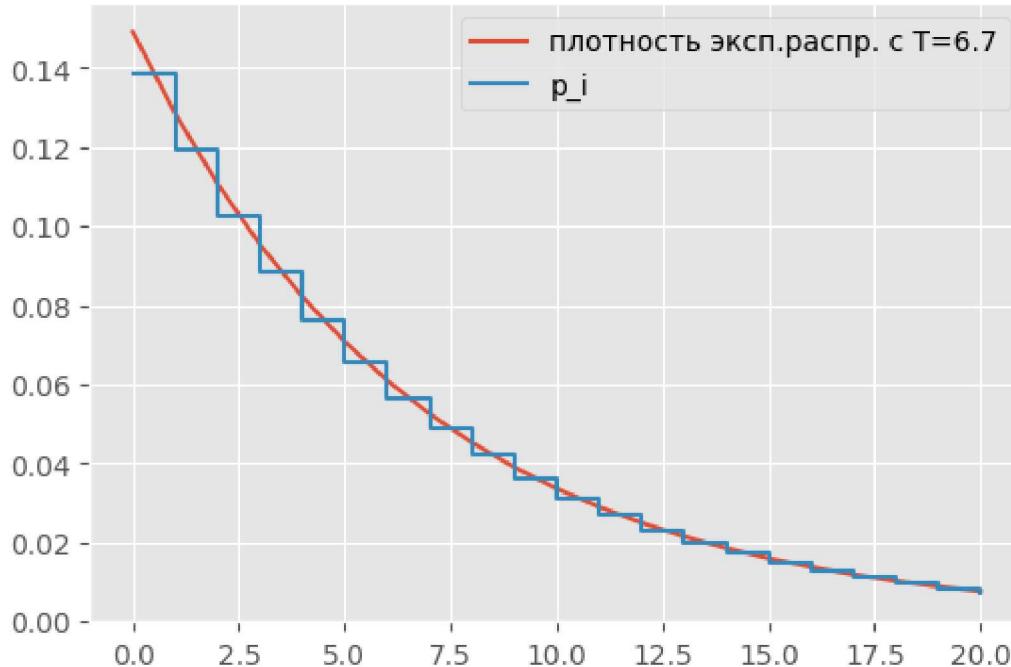
$$M(g(X)) = \sum_{x=1}^{+\infty} p(x) \cdot bonus(x) dx$$

Вопрос: как посчитать $p(x)$, $x=1,2,\dots$

Можно воспользоваться эмпирическим законом распределения, а лучше восстановленным теоретическим:

$$p(X = i) = F_X(i) - F_X(i - 1) \approx f_X(i - 0.5) \cdot (i - (i - 1)) = f_X(i - 0.5)$$

```
# нарисуем плотность
xbins = np.arange(0, 21)
plt.figure(figsize=(6, 4))
plt.plot(xbins, exp_val.pdf(xbins), label=f"плотность эксп.распр. с T={Tmean}")
x_i = np.array([xbins[(i+1) // 2] for i in range(0, 2 * len(xbins) - 1)])
y_i = np.array([exp_val.pdf(xbins[i // 2] + 0.5) for i in range(0, 2 * len(xbins) - 1)])
plt.plot(x_i, y_i, label='p_i')
plt.legend()
plt.show()
```



видим, что сумма площадей под ступеньками будет примерно равна площади под кривой. Т.е. мы дискретизируем СВ X.

```

bonus = lambda x: 500 + 1000 / x

# подсчитаем bonus_i, i=1,2,3,...80
days = np.arange(1, 81)
bonus_i = np.array([bonus(x) for x in days])
bonus_i[:10]

array([1500.    , 1000.    , 833.333, 750.    , 700.    , 666.667,
       642.857, 625.    , 611.111, 600.    ])

```

```

# подсчитаем p_i и мат.ожидание бонуса с использованием плотности распределения
p_i = exp_val.pdf(days - 0.5)
m_bonus = bonus_i @ p_i
print(f"M(bonus) = {m_bonus}")

M(bonus) = 817.2785212847335

```

```

# подсчитаем p_i и мат.ожидание бонуса с использованием функции распределения
p_i = exp_val.cdf(days) - exp_val.cdf(days - 1)
m_bonus = bonus_i @ p_i
print(f"M(bonus) = {m_bonus}")

M(bonus) = 818.0373262613218

```



ЗАДАНИЕ.

Пусть СВ "Результат подключения клиента по заявке" подчиняется закону Бернулли с параметром $p=0.25$, т.е.:

$$p(x) = \begin{cases} 0.75 & x = 0 \\ 0.25 & x = 1(\text{подключен}) \end{cases}$$

Найдите мат.ожидание бонуса, полученного отделом за подключение заявок за год, если бонус за каждую подключенную заявку составлял (в среднем) 800 руб и за год было подано 1000 заявок.

ВАРИАНТЫ ОТВЕТОВ:

1. 100 т.р.
2. 200 т.р.
3. 400 т.р.
4. 800 т.р.

▼ Свойства математического ожидания

Свойство 1. МО суммы равно сумме МО:

$$M[X + Y] = M[X] + M[Y]$$

Свойство 2. Постоянный множитель для СВ может быть вынесен из-под знака МО:

$$M[k \cdot X] = k \cdot M[X]$$

Таким образом можно считать оператор вычисления МО линейным (МО линейной комбинации СВ равен линейной комбинации МО):

$$M[a \cdot X + b \cdot Y] = a \cdot M[X] + b \cdot M[Y]$$

Свойство 3. МО произведения двух независимых СВ равно произведению их МО:

$$M[X \cdot Y] = M[X] \cdot M[Y]$$

▼ Начальные и центральные моменты

Какие характеристики законов распределения СВ являются для нас особенно ценными? С частью из них мы уже познакомились. Это:

- среднее значение СВ (m, μ, loc);
- СКО или среднеквадратическое отклонение СВ от его среднего (σ, scale) или корень квадратный из дисперсии (D, Var).

Но это только два момента СВ. Давайте разберемся а что такое моменты для СВ и какие моменты нам особенно полезны.

▼ Начальный момент k -го порядка

Опр. Начальным моментом k -го порядка для СВ X будем называть математическое ожидание для X^k .

Для ДСВ :

$$\nu_k(X) = \sum_{\forall x_i} p_i \cdot x_i^k$$

Здесь $p_i = p(X = x_i)$

или, если нам доступна ГС, то просто находим среднее по всем значениям из ГС:

$$\nu_k(X) = \frac{1}{|\Gamma C|} \sum_{\forall x_i \in \Gamma C} x_i^k$$

Последняя формула теоретически подходит и для НСВ, но как мы понимаем, с практической точки зрения бессмысленна (Почему?).

В общем случае, для НСВ X момент или среднее значение СВ X или какой-то функции от СВ $g(x)$ находят с помощью интеграла и плотности распределения СВ $f_X(x)$:

$$\nu_k(X) = \int_{-\infty}^{+\infty} f_X(x) \cdot x^k dx$$

▼ ПРИМЕР. Анализ времени подключения заявок

Вернемся к нашему примеру со СВ X ="Время подключения клиента по заявке, поданной в 2023 году". Оказывается, мы уже находили начальные моменты для X

```
import numpy as np
from scipy import integrate
import scipy.stats as stat
# найдем начальные моменты 1-го и 2-го порядков (мат.ожидание от X и от X**2)

# зададим подинтегральные функции
Tmean = 6.7
exp_val = stat.expon(scale=Tmean)
f_X = lambda x: exp_val.pdf(x) * x
f_X2 = lambda x: exp_val.pdf(x) * x**2

# найдем M(X) - выполним интегрирование с помощью пакета
M_X = integrate.quad(f_X, -np.inf, +np.inf)[0]
M_X2 = integrate.quad(f_X2, -np.inf, +np.inf)[0]

# посмотрим результат
print(f"Начальный момент 1-го порядка: M(X) = {round(M_X, 3)}")
print(f"Начальный момент 2-го порядка: M(X**2) = {round(M_X2, 3)}")

Начальный момент 1-го порядка: M(X) = 6.7
Начальный момент 2-го порядка: M(X**2) = 89.78
```

Если начальный момент 1-го порядка - это просто среднее значение (МО) СВ X , то что показывает начальный момент 2-го порядка?

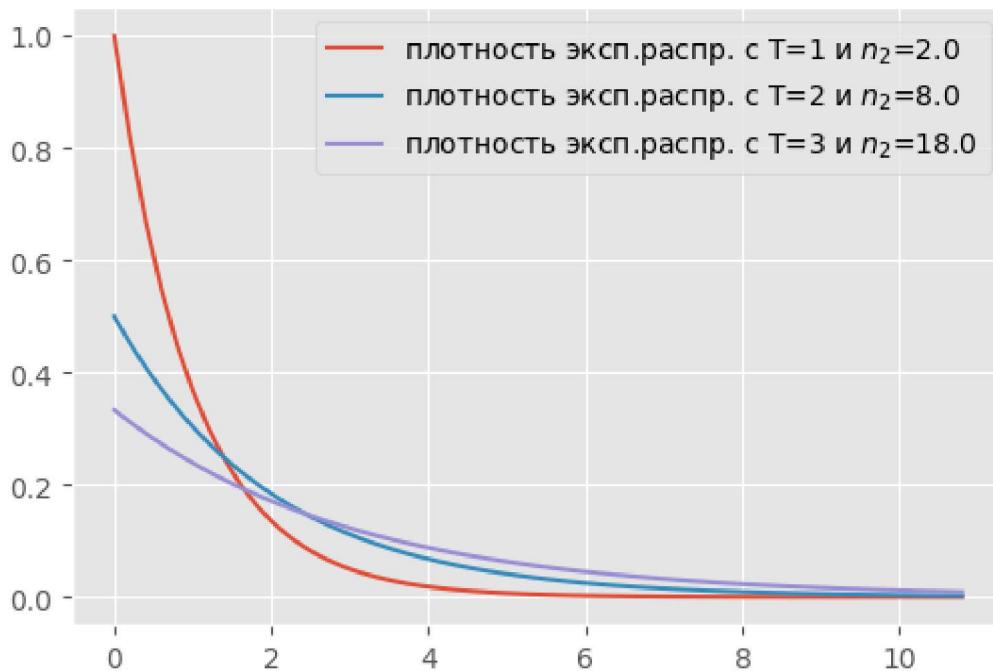
Для тех, что знает физику - это момент инерции силы, распределенной по оси ОХ относительно центра.

Проще говоря, он показывает насколько далеко от нуля СВ X может принимать значения. Чем больше этот момент, тем больше "хвосты" распределения отстоят от начала координат.

Сравним моменты 2-го порядка для трех экспоненциальных распределений

```
x_means = [1, 2, 3]
f_X2 = lambda x: exp_val.pdf(x) * x**2
```

```
# нарисуем плотности экспоненциальных распределений
xbins = np.arange(0, 11, 0.2)
plt.figure(figsize=(6, 4))
for i, Xmean in enumerate(x_means):
    exp_val = stat.expon(scale=Xmean)
    M_X2 = integrate.quad(f_X2, -np.inf, +np.inf)[0]
    plt.plot(xbins, exp_val.pdf(xbins), label=f"плотность эксп.распр. с T={x_means[i]} и")
plt.legend()
plt.show()
```



Видим, что начальный момент 2-го порядка очень быстро (квадратично) растет с ростом среднего значения СВ X. А если быть совсем точным, то для СВ X, распределенной по экспоненциальному закону, верно следующее соотношение:

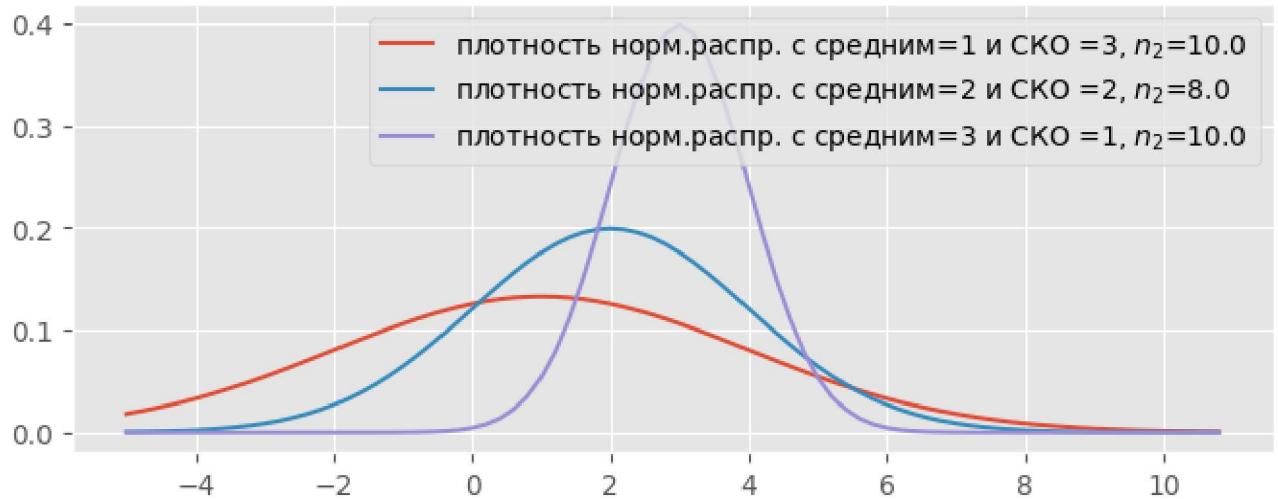
$$\nu_2(X) = 2 \cdot (M(X))^2$$

Для нормально распределенных СВ такой зависимости нет

```
x_means = [1, 2, 3]
sigma_list = [3, 2, 1]
f_X2 = lambda x: norm_val.pdf(x) * x**2

# нарисуем плотности экспоненциальных распределений
xbins = np.arange(-5, 11, 0.2)
plt.figure(figsize=(8, 3))
for i in range(len(x_means)):
    Xmean = x_means[i]
    sigma = sigma_list[i]
    norm_val = stat.norm(loc=Xmean, scale=sigma)
    M_X2 = round(integrate.quad(f_X2, -np.inf, +np.inf)[0], 2)
```

```
label=f"плотность норм.распр. с средним={Xmean} и СКО ={sigma}, $n_2$={M_X2}"  
# label = f"плотность норм.распр. {chr(65+i)}"  
plt.plot(xbins, norm_val.pdf(xbins), label=label)  
  
plt.legend()  
plt.show()
```



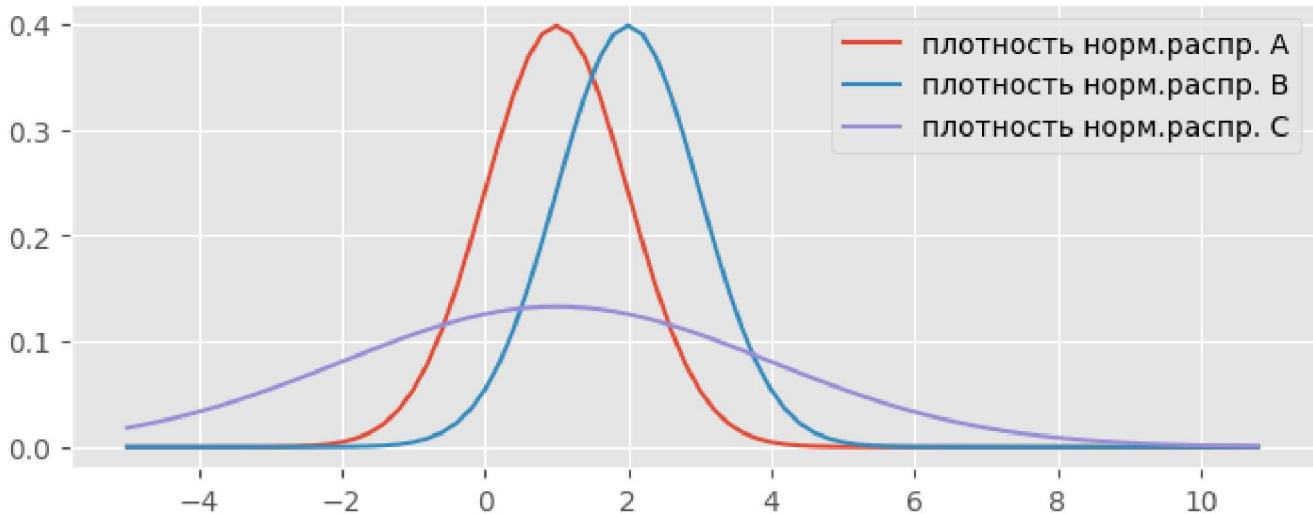
Сейчас мы видим, что нам важнее получить информацию об отклонении хвостов не от 0, а от среднего значения СВ X . Для этого и существуют центральные моменты. Они хорошо характеризуют СВ с симметричным относительно среднего распределением.



▼ ЗАДАНИЕ

Для выведенных ниже плотностей распределения расставьте их в порядке возрастания начального момента 2-го порядка:

1. A, B, C
2. A, C, B
3. C, A, B
4. C, B, A



▼ Центральный момент k-го порядка

Опр. Центральным моментом k-го порядка для СВ X будем называть математическое ожидание для $(X - M(X))^k$.

Для ДСВ :

$$\mu_k(X) = M[(x - m_X)^k] = \sum_{\forall x_i} p_i \cdot (x_i - m_X)^k$$

Здесь $p_i = p(X = x_i)$, $m_X = M(X)$

или, если нам доступна ГС, то просто находим среднее по всем значениям из ГС:

$$\mu_k(X) = \frac{1}{|\Gamma C|} \sum_{\forall x_i \in \Gamma C} (x_i - m_X)^k$$

для НСВ X :

$$\mu_k(X) = M[(x - m_X)^k] = \int_{-\infty}^{+\infty} f_X(x) \cdot (x - m_X)^k dx$$

Центральный момент 1-го порядка нам неинтересен, так как в соответствии со свойством линейности МО имеем:

$$\mu_1(X) = M[X - m_X] = M[X] - M[m_X] = m_X - m_X = 0$$

Центральный момент 2-го порядка называют **дисперсией**. Она также, как и начальный момент 2-го порядка характеризует насколько сильно хвосты распределения отстоят от среднего значения СВ:

$$\mu_2(X) = D_X = \sigma_X^2 = M[(X - m_X)^2]$$

Воспользовавшись свойствами МО можно вывести:

$$D_X = M[(X - m_X)^2] = M[X^2 + m_X^2 - 2 \cdot X \cdot m_X] = M[X^2] + m_X^2 - 2 \cdot m_X \cdot M[X] \\ - m_X^2 = \nu_2 - \nu_1^2$$

Проверим это соотношение на примере экспоненциального распределения

```
# зададим подинтегральные функции
exp_val = stat.expon(scale=3)
f_X = lambda x: exp_val.pdf(x) * x
# найдем M(X) - выполним интегрирование с помощью пакета
M_X = integrate.quad(f_X, -np.inf, +np.inf)[0]

f_X2 = lambda x: exp_val.pdf(x) * x**2
f_X_2 = lambda x: exp_val.pdf(x) * (x - M_X)**2

M_X2 = integrate.quad(f_X2, -np.inf, +np.inf)[0]
D_X = integrate.quad(f_X_2, -np.inf, +np.inf)[0]

print(f"Мат.ожидание X M(X) = {round(M_X, 3)}")
print(f"Мат.ожидание X**2 M(X**2) = {M_X2}")
print(f"M(X**2) - M(X)**2 = {round(M_X2 - M_X**2, 3)}")
print(f"ДИСПЕРСИЯ D(X) = M[(X - M(X))**2] = {round(D_X, 3)}")
print(f"СКО = Корень из ДИСПЕРСИИ = {round(np.sqrt(M_X2 - M_X**2), 3)}")
```

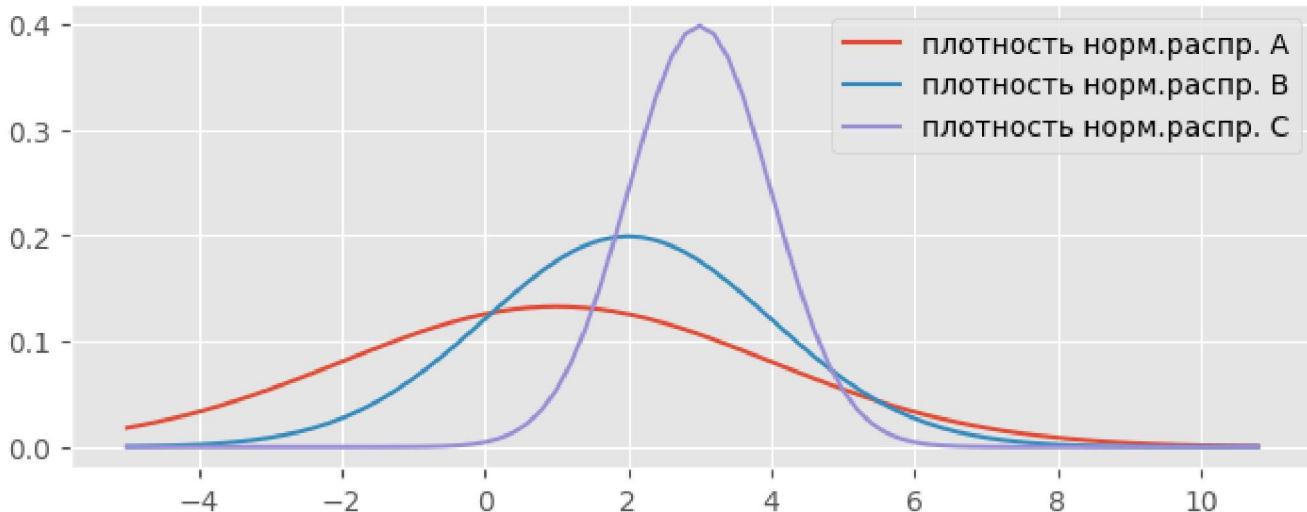
Мат.ожидание X M(X) = 3.0
 Мат.ожидание X**2 M(X**2) = 18.0
 $M(X^{**2}) - M(X)^{**2} = 9.0$
 ДИСПЕРСИЯ D(X) = $M[(X - M(X))^{**2}] = 9.0$
 СКО = Корень из ДИСПЕРСИИ = 3.0



▼ ЗАДАНИЕ.

Для выведенных ниже плотностей распределения расставьте их в порядке возрастания дисперсии:

1. A, B, C
2. A, C, B
3. C, A, B
4. C, B, A



▼ Статистики как функции векторной СВ, выборочные оценки

- На практике мы всегда работаем с выборками из ГС.
- На основе этих выборок мы строили эмпирические законы распределения. Для того, чтобы построить эмпирический закон распределения (функцию вероятности, гистограмму распределения) необходимо большое количество наблюдений (для устойчивости оценок p_i или h_i). Не всегда нам доступно большое кол-во наблюдений и тем более не всегда эти наблюдения охватывают весь спектр возможных значений СВ.
- с практической точки зрения гораздо продуктивнее сделать предположение о виде распределения (или смеси распределений), а затем оценить параметры распределения (среднее, дисперсию, ...).

Таким образом, мы приходим к необходимости изучения выборочных оценок параметров, которые сами являются СВ. И эти СВ подчиняются определенным законам распределения. Так что мы должны научиться строить эти законы и определять свойства этих оценок.

Опр. Выборкой из ГС размера n будем называть набор значений СВ X , который мы наблюдали или можем наблюдать при n экспериментах.

Обозначать конкретную выборку из ГС СВ X размером n будем набором (x_1, \dots, x_n) . Обозначать мыслимую выборку из ГС СВ X размером n (векторную СВ) будем большими буквами (X_1, X_2, \dots, X_n) .

Опр. Статистика – числовая функция от выборки:

$$f(X_1, X_2, \dots, X_n)$$

Мы уже знаем примеры таких статистик как:

- среднее (выборочное):

$$x_{\text{cp}}(n) = \bar{x}(n) = f(X_1, X_2, \dots, X_n) = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Аналогично выборочному среднему мы можем вычислять другие выборочные моменты k-го порядка:

- выборочный начальный момент 2-го порядка:

$$\bar{x^2}(n) = f(X_1, X_2, \dots, X_n) = \frac{X_1^2 + X_2^2 + \dots + X_n^2}{n}$$

- выборочная дисперсия:

$$S_X^2(n) = \frac{(X_1 - m_X)^2 + (X_2 - m_X)^2 + \dots + (X_n - m_X)^2}{n}$$